The Dissertation Committee for Soojin Kim Certifies that this is the approved version of the following dissertation:

An Automated Test Assembly for Unidimensional IRT Tests Containing Cognitive Diagnostic Elements

Committee:

Hua-Hua Chang, Supervisor

Barbara G. Dodd

William R. Koch

S. Natasha Beretvas

Charles N. Friedman

An Automated Test Assembly for Unidimensional IRT Tests Containing Cognitive Diagnostic Elements

by

Soojin Kim, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2004

Dedication

To my husband Jung Su

and my parents.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Hua-Hua Chang for his advice and support throughout my doctoral work. He was such a great advisor and mentor, who always made time to listen and provide prudent advice and encouragement when I needed them. I also would like to thank to my committee members, Dr. Barbara Dodd, Dr. Bill Koch, Dr. Tasha Beretvas, and Dr. Charles Friedman for providing me guidance along the way.

I also want to thank past and current members of Dr. Chang's research group for all of their help in the laboratory. Thanks also go to Meghan Wills, Ying Cheng and Pei-hua Chen for encouraging me and supporting me every minute.

I would like to thank my parents, Dae Hee Kim and Young-ha Kim and all my family for supporting me and encouraging me. Last and not least, I thank my lovely husband for his endless love and care.

An Automated Test Assembly for Unidimensional IRT Tests Containing Cognitive Diagnostic Elements

Publication No._____

Soojin Kim, Ph.D.

The University of Texas at Austin, 2004

Supervisor: Hua-Hua Chang

Large-scale assessments are typically administered numerous times per year using parallel test forms. The traditional methods of constructing parallel test forms are based on manually selecting items for given test specifications such as content balancing. This methods are cumbersome, time consuming, and inefficient. To overcome these problems, an automated test assembly has been used successfully in test construction to assemble conventional IRT tests (van der Linden, 1994). However, these conventional large-scale assessments only provide a single summary score that indicates the overall performance level or achievement level of a student in a single learning area. For assessments to be more effective, tests should provide useful diagnostic information in addition to single overall scores. One approach is using a Cognitive Diagnosis modeling. The purpose of this research is to develop an algorithm for generating information-rich tests by combining Cognitive Diagnosis with the traditional IRT approach that not only

produce a single score to measure an examinee's ability level but also provide diagnostic information. This study describes a new method of automated test assembly, which incorporates diagnostic techniques with existing IRT-based testing assembly methods.

The purpose of Cognitive Diagnosis modeling is to provide useful information by estimating individual knowledge states by assessing whether an examinee has mastered specific attributes measured by the test (Embretson, 1990; DiBello, Stout, & Rousses, 1995; Tatsuoka, 1995). Attributes are skills or cognitive processes that are required to perform correctly on a particular item. If standardized testing could incorporate assessments of the various attributes constituting the item, then students, parents, and teachers would be able to see where a student stands with respect to mastering the item. Such information could be used to guide the learner toward areas requiring more study. Helping students to identify their intellectual strengths and weaknesses is more informative and instructive than simply giving them a single score that represents their overall ability. By being able to assess where they stand in regard to the attributes that compose an item, students can plan a more effective learning path to be desired proficiency levels.

Even though Cognitive Diagnosis has attracted considerable attention from researchers, few studies have described how to assemble a test that conforms to given cognitive criteria. If such a test could be assembled, it would provide more specific identification of the areas where students needs to improve their skills. Also, it would provide diagnostic feedback to teachers, who could then address the specific needs of individual students. In this way, the test becomes an active tool in the educational process rather than just a passive score report.

The proposed automated test assembly method and its corresponding computer algorithm will be developed to construct tests automatically from a given item bank while assuring the tests conform to specifications from both conventional IRT scaling and the Cognitive Diagnostic aspects. The method employs the commonly used Zero-One (0/1) Linear Programming Method. This study describes a new method of automated test assembly, which incorporates diagnostic techniques with existing IRT-based testing assembly methods using Maxmin, Minimax, and Maximum Information Methods. A major goal of this research is to identify a set of the most reasonable constraints in Cognitive Diagnosis and to integrate those new constraints into traditional IRT scaling.

Most traditional test assembly methods tend to select best test items to form a test under given test specifications, such as content balancing, item difficulties, item formats, reliabilities, test length, and many more (van der Linden, 1998). For this research, a component to deal with Cognitive Diagnosis is added to the current existing automated test assembly method based on IRT. The research described in this dissertation sought to apply and improve available technologies to automate this task and thereby contribute to a new area of educational research. By implementing the Cognitive Diagnostic approach within the traditional standardized test assembly methods, testing specialists will find that using the algorithm introduced in this dissertation might prove useful to test development.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

One of the best ways to learn about students is to give them tests: for example, achievement tests, personality tests, aptitude tests, and many more. The demand for such information has led to extensive standardized testing, so it is important that we be able to transform standardized test results into general skill-level profiles, that can be used to guide teaching and learning processes. Typically, however, large-scale assessments provide only a single summary score that indicates the overall performance or achievement level of a student in one learning area. For assessment to be more effective, tests should provide useful diagnostic information in addition to single overall scores.

Information on how students are performing is especially needed these days to meet provisions of the No Child Left Behind Act of 2001 (NCLB), which mandates a common standard for all students, schools, and states. In particular, the NCLB requires all students to perform at some level of *proficiency* by the 2013-2014 school year. What that level of *proficiency* is, however, varies from state to state, because the U.S. Department of Education (USED) requires the states to define the term for themselves. To improve student proficiency, each state must also develop a good measure of *proficiency* and a workable procedure for feedback.

In the context of NCLB, while educators appreciate the need for diagnostics, few agree on what diagnostics are, and difficulties in defining diagnostics have paralyzed decision-making concerning the use of test scores from assessments.

Because any single score intended to assess a target skill is unlikely to yield a diagnosis that is rich and meaningful for students and parents-- as well as for teachers--assessment designs are needed by which assessments may provide more than single scores.

One approach to this problem is Cognitive Diagnosis modeling--founded on Item Response Theory (IRT) modeling--whose purpose is to provide useful information by estimating individual knowledge states by assessing whether an examinee has mastered specific attributes measured by the test. Attributes are skills or cognitive processes that are required to perform correctly on a particular learning item (Chipman, Nichols, & Brennan, 1995).

Attributes can be illustrated as a Q-matrix, the description of which items measure which attributes (Tatsuoka, 1983). The Q-matrix is a *K* x *n* matrix containing ones and zeros, where *K* indicates the number of attributes we wish to assess and *n* indicates the number of items on the test. Each cell of the Q-matrix, q_{ik}, takes a value of 1 if mastering skill *k* is required to solve item i, and 0 otherwise (Tatsuoka, 1983). Then, the Q-matirx can be translated into a form that can be compared to individual observed item response patterns. This is achieved by identifying a α vector for each examinee: $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$, with the *k*th element, α_k, indicating whether the examinee masters attribute *k* or not.

If standardized testing could incorporate assessments of the various attributes constituting the learning item, then students, parents, and teachers would be able to see where a student stands with respect to mastering the item. Such information

could be used to guide the learner toward areas requiring more study (Campione & Brown, 1990). Moreover, the assessment could be linked to specific classroom activities. Helping students to identify their intellectual strengths and weaknesses is more informative and instructive than simply giving them a single score that represents their overall ability. By being able to assess where they stand in regard to the attributes that are assessed by an item, students are able to plan a more effective learning path to desired proficiency levels.

Even though Cognitive Diagnosis has attracted considerable attention from researchers, few studies have described how to assemble a test that conforms to given cognitive criteria. If such a test could be assembled, it would provide more specific identification of the areas where students needs to improve their skills and would provide diagnostic feedback to teachers, who could then address the specific needs of individual students (Embretson, 1990). In this way, the test becomes an active tool in the educational process rather than just a passive score report.

The purpose of this dissertation research is, therefore, to develop an algorithm for generating such information-rich tests by combining Cognitive Diagnosis with the traditional unidimensional IRT approach that produces a single score to measure an examinee's ability level.

Traditional test assembly methods provide for considering the test specifications that aid item selection, such as content balancing, item difficulties, item formats, reliabilities, test length, and many more (van der Linden, 1998). These conventional techniques for manually selecting items are time-consuming and

cumbersome. To overcome the problems of the conventional methods, an automated test assembly has been used successfully in test construction to assemble traditional unidimensional IRT tests.

This dissertation, however, describes a new automated test assembly method, one that incorporates diagnostic techniques with existing unidimensional IRT testing methods. The automated test assembly method, which is a computer algorithm, are developed to construct tests automatically from a given item bank while assuring the tests conform to specifications both from conventional unidimensional IRT scaling and from the Cognitive Diagnostic approach. The method employs the commonly used Zero-One (0/1) Linear Programming Model.

A number of Linear Programming methods have been developed to solve various psychometric problems (for example, Adema & van der Linden, 1989; Boekkooi-Timminga, 1987, 1990; de Gruijter, 1990; Theunissen, 1985, 1986; van der Linden & Boekkoi-Timminga, 1988, 1989). Among those, Theunissen (1985) was the first to present a 0/1 Linear Programming Method for test construction with a target information function. In the field of test assembly, the 0/1 Linear Programming Method is properly defined as a combinatorial optimization process. The combinatorial optimization involved in the automated test assembly approach uses an optimal item pool that consists of a maximal number of combinations of items that (1) meet all content specifications for the test and (2) are most informative of a series of ability levels reflecting the shape of the distribution of the ability estimates for the population of the examinee (van der Linden, 1998).

For this dissertation, an automated test assembly method is established to develop a test based on the Cognitive Diagnosis and IRT. First, the objective function of the automated test assembly method is optimized (for example, by maximizing test information or minimizing the sum of the positive deviations from the target test information). Second, constraints are formulated according to test specifications in conventional unidimensional IRT tests (such as test length and contents). Finally, a set of new constraints is added to the conventional IRT constraints in order to express the Cognitive Diagnostic aspect of the procedure (such as an assembled test Q-matrix and the information related to the discriminant).

A major goal of this research was to identify some constraints in Cognitive Diagnosis and to integrate those new constraints into traditional unidimensional IRT scaling. By using the real responses from samples of 2,000 students, an item pool of 542 items was constructed. By using this item pool and test specifications from a real large-scale educational assessment (3^{rd} grade math exams from the TASS assessment of the Texas Education Agency), tests are automatically generated using a commercial software GAMS (Boisvert, Howe, and Kahaner, 1985) with CPLEX solver (ILOG, 2003). Then, simulation studies were conducted to compare the results.

This dissertation includes six topics important for understanding the implementation of this research:

- a brief description of traditional Item Response Theory
- a description of the newly developed Cognitive Diagnosis Theory that includes two groundwork models of Fisher's LLTM and Tatsuoka's Rule Space Methodology as well as a Unified Model and a Fusion Model
- a description of traditional automated test assembly methods
- a description of the 0/1 Linear Programming Method that is used for automated test assembly
- a description of new automated test assembly for Cognitive Diagnosis
- a presentation of results from a simulation study comparing three proposed automated test assembly methods.

Combining the strengths of the conventional testing framework and the new cognitive diagnostic framework, this new method will benefit many fields in educational and psychological testing. So, while Cognitive Diagnostic assessments can help both learners and educators by giving them the means to diagnose learners' knowledge states correctly, time and effort are wasted when tests must be assembled manually. The research described in this dissertation sought to apply and improve available technologies to automate this task and thereby contribute to a new area of educational research. By implementing the Cognitive Diagnostic approach within the traditional standardized test assembly methods, testing specialists will find that

using the algorithm introduced in this dissertation might prove useful to test development.

CHAPTER 2: LITERATURE REVIEW

This chapter first reviews the assumptions and characteristics of Item Response Theory (IRT) and then describes a specific application of IRT, Cognitive Diagnosis modeling. Two key concepts are explained: attributes and the cognitive processes of problem solving. The chapter then examines several models—specifically Fischer's Linear Logistic Trait Model (LLTM), Tatsuoka's Rule Space Methodology, the Unified Model, and the Fusion Model. Next is a discussion of the automated test assembly method, which makes use of binary programming. The chapter concludes with a description of the Zero-One (0/1) Linear Programming Model for item selection using item information as an objective function.

2.1 ITEM RESPONSE THEORY

Item Response Theory (IRT), also known as latent trait theory, has been applied more and more frequently as a foundation theory in educational and psychological measurement research. IRT models use a mathematical function to explain a relationship between observable performance and certain unobservable traits or abilities (Rogers, Swaminathan, and Hambleton, 1991). The purpose of IRT is to provide a foundation for making estimates (or predictions) about abilities (or traits) measured by a test (Hambleton & Swaminathan, 1985). One of the most important characteristics of IRT is that the item, as opposed to the whole test, is the

unit of measure used to obtain ability scores on the same scale, regardless of differences found when administering items across examinees (Wainer & Mislevy, 2000). While IRT is a well-developed and robust theory, the assumptions of IRT models regarding test data are known to be difficult to meet.

2.1.1 Assumptions

One important assumption made under the most common IRT models is that of unidimensionality, meaning that test items measure only one kind of ability (Hambleton & Swaminathan, 1985). According to this assumption, only one dominant factor affects test performance. This dominant factor, also called a latent trait (θ), can be explained by the performance of an individual on a set of test items (Rogers et al., 1991).

Another important assumption is that of local independence, meaning that, given that the abilities influencing the test performance are held constant, the responses of an examinee to any pair of items are statistically independent (Hambleton & Swaminathan, 1985). Use of the IRT model requires that the local independence assumption be met, because the response pattern probability is achieved simply by multiplying the individual item probabilities (Embretson & Reise, 2000). This assumption can be used as an important concept for test information, which will be explained in Section 2.1.3.

Third, IRT assumes that a monotonically increasing function represents the relationship between true trait levels and the performances of individuals indicating those trait levels (Hambleton & Swaminathan, 1985). For a given examinee, as the trait level or ability level increases, the probability of responding to the item correctly increases as well (Rogers et al., 1991).

2.1.2 Models

Based on these assumptions, many kinds of IRT models have been developed, models that can accommodate both dichotomous and polytomous item-response possibilities. Among these possible IRT probability models, a few of the most common models are briefly discussed. The IRT models the probability of a correct response to a test item as a function of one or more parameters of the item (typically designated *a*, *b*, and *c*) and the latent trait level of examinee *j*, typically designated θ_j (Rogers et al., 1991). One way to interpret the expression $P_{ij}\left(Y_{ij}=1|\theta_j\right)$ is to think of it as the proportion of individuals, each with ability θ, who correctly answer item *i*. When Y_{ij} = 1, the answer of item *i* is correct, and Y_{ij} = 0 if the answer is incorrect. Accordingly, this logistic model relates the level of the item parameter and the person parameters to the probability of responding correctly.

In this section, the three most commonly used dichotomous IRT models are described: the three-parameter logistic model (3PL), the two-parameter logistic model (2PL), and the one-parameter logistic model (1PL). The 3PL, 2PL, and 1PL models

are referred to as dichotomous because they may be applied to tests with multiple-choice items and short constructed-response items that are scored either correct (scored as 1) or incorrect (scored as 0).

The 3PL Model

The 3PL model is given as follows (Birnbaum, 1968):

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j\right)=c_i+(1+c_i)\frac{e^{Da_i\left(\theta_j-b_i\right)}}{1+e^{Da_i\left(\theta_j-b_i\right)}} \quad (1)$$

where $P_{ij}\left(Y_{ij}=1\middle|\theta_j\right)$ is the probability that an examinee with ability θ_j answers test item i correctly. The threshold parameter of item i, denoted as b_i, is the item difficulty parameter. Because in IRT items and persons are on the same scale, it can be said that Person A's trait level is almost the same as Item 1's difficulty, or Item 1 is almost as hard as Person A's trait level (Embretson & Reise, 2000).

The slope parameter of item j, denoted as a_i, is the item discrimination parameter that characterizes the sensitivity to proficiency (Hambleton & Swaminathan, 1985). The value of the item discrimination parameter, a_i, is said to be proportional to the slope of the probability function at the location of b_i on the ability axis (Rogers et al., 1991). Therefore, this discrimination parameter controls how steep the ICC lies. Thus, using this parameter, students can be distinguished with trait levels above and below the rising slope of the ICC.

The lower asymptote parameter of item *i*, denoted as c_i, is what is termed the "guessing" or “pseudo-chance level” parameter (Rogers et al., 1991). This c_i parameter reflects the chance that a student who has very low proficiency will nevertheless select the correct option (Hambleton & Swaminathan, 1985). Among the three models, the three-parameter model is said to be the most realistic in that it acknowledges the chance correct response through c_i.

The 2PL Model

The defining equation for the 2PL model is the same as that for the 3PL model except that in the 2PL model the c_i parameter is fixed as zero. The 2PL model is shown below:

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j\right)=\frac{e^{Da_i\left(\theta_j-b_i\right)}}{1+e^{Da_i\left(\theta_j-b_i\right)}} \tag{2}$$

for $i = 1, 2, \ldots, n$, and where D is a scaling constant. This model contains both the item difficulty parameter b_i and the item discrimination parameter a_i (Birnbaum, 1968). Item discrimination depicts the item’s capability in discriminating among examinees of different ability levels. The 2PL model, given an ability level θ_j, gives the probability of getting a correct response to item *i* as shown in equation (2) (Embretson & Reise, 2000).

The 1PL Model

The 1PL model, also called the Rasch model (Rasch, 1960), is as follows:

$$P_{ij}\left(Y_{ij}=1\middle|\theta_j\right)=\frac{e^{(\theta_j-b_i)}}{1+e^{(\theta_j-b_i)}} \tag{3}$$

for i = 1, 2, …, n. This model indicates the probability of obtaining a correct response to an item given a certain level of ability and item difficulty (Hambleton & Swaminathan, 1985). The model contains only the item difficulty parameter b_i for each item i. Other parameters, such as the item discrimination parameter a_i and the item guessing parameter c_i, are set to zero.

In contrast to dichotomous items, some items are scored on a multipoint scale, with scores ranging from 0–3 or 0–6, for example (Embretson & Reise, 2000). Several polytomous IRT models have been developed, such as the Graded Response Model (Samejima, 1997), Partial Credit Model (Master, 1982), Generalized Partial Credit Model (Muraki, 1992), and many more. For more a detailed explanation of the polytomous models, please refer to the book of Embretson and Reise (2000).

2.1.3 Item and Test Information

Certain features of the IRT models are essential for test assembly. The above parameters are used especially as practical ingredients for the test assembly that will be described in later sections.

One of the most important features of the IRT models is embodied in the concept of item information. An item information curve is said to be transformed from an item-response curve from any kind of dichotomous IRT model or the category-response curves from a polytomous IRT model (Embretson & Reise, 2000). This item information curve specifies how much Fisher information each item contains at all points on the latent-trait continuum.

For dichotomous IRT models, the item information function (IIF) is estimated as follows:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \quad (4)$$

where $P_i(\theta)$ is the probability of correctly responding to item *i* given ability θ, and $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ (Lord, 1980).

One of the advantages of IIF curves is that they can be added to specify the shape of the curve for the test information function (TIF) (Hambleton & Swaminathan, 1985). Under the local independence assumption, the total amount of information for a test can be readily verified. This curve is one of the most

important characteristics when the automated test assembly is used, for reasons to be addressed shortly. As simply the sum of IIF curves, the TIF curve is expressed as follows:

$$TI_t(\theta) = \sum_{i=1}^{n} I_i(\theta). \qquad (5)$$

This test information is critically crucial in determining how well a test performs because the degree of item information is the reciprocal of the standard error of measurement, shown in Equation 6 (Lord, 1977).

$$SE\ (\theta) = \frac{1}{\sqrt{TI(\theta)}} \qquad (6)$$

Items with low standard errors provide greater information, and items with high standard errors provide less information (Lord, 1977). Greater item information provides the test constructor with greater precision in measurement and helps in the selection of items to include in a test. Item information and test information can be used in basic test design.

The sections that follow describe a different approach to test measurement, one using the Cognitive Diagnostic Model, and discuss how the new Cognitive Diagnostic assessment approach can be combined with the traditional unidimensional IRT approaches for test assembly purposes.

2.2 COGNITIVE DIAGNOSIS THEORY

Increasingly, psychometricians are showing an interest in the new theory, Cognitive Diagnosis, because it offers an important addition to IRT-based test measurement techniques. Like traditional IRT models, IRT-based Cognitive Diagnosis models also define the probability of examinee *j*'s response to item *i*, given the examinee's ability parameters and the item parameters. More importantly, however, a cognitive-diagnosis-model-based assessment is a skills-level "formative" assessment and a tool that can aid the teaching and learning processes (Embertson, 1990).

The purpose of the Cognitive Diagnostic assessment is both to evaluate examinees cognitively and to evaluate test items cognitively (Hartz, Roussos, and Stout, 2002). Instead of assigning a single ability estimate to each examinee, as is done in typical IRT-based summative assessments, the formative assessments based on the Cognitive Diagnosis model divide the latent space multidimensionality into more refined, often discrete or dichotomous, cognitive skills or latent attributes (DiBello, Stout, & Roussos, 1995). In addition, Cognitive Diagnosis evaluates examinees with respect to their level of competence in each attribute. This evaluation can be done by giving individual feedback to examinee using the attributes measured by the assessment. An *attribute* is identified as a "task, subtask, cognitive process, or skill" involved in the assessment (Tatuoka, 1995, p.330).

Cognitive Diagnosis modeling has two major strengths (Hartz et al, 2002):

1. It determines attribute mastery or non-mastery profiles of the examinees taking a test.
2. It evaluates the test and its items in terms of their effectiveness in measuring the individual attributes.

These two advantages work together to make the models more efficient. Hartz (2002) mentions several other IRT-based cognitive diagnostic models that have been proposed during the development stage of cognitive diagnosis research; however, these models have only a few of the obvious requirements for effective cognitive diagnosis, namely, that they cognitively evaluate examinees, cognitively evaluate items, or incorporate statistically identifiable parameters. A table of the fourteen models reviewed by Hartz is provided in Appendix A; for a more detailed description of the models, refer to Hartz (2002).

Among the fourteen models, two earlier models represent the groundwork for the development of the Cognitive Diagnosis models. They are Fisher's Linear Logistic Trait Model (LLTM) (Fischer, 1973) and Tatsuoka & Tatsuoka's Rule Space Methodology (Tatsuoka & Tasuoka, 1982). Fisher's LLTM is an item-based Cognitive Diagnosis model, and Tatsuoka & Tatsuoka's Rule Space Methodology is examinee-based Cognitive Diagnosis modeling. Based on these two models, a Unified Model (DeBello et al., 1995) was developed, and then a Fusion Model (Hartz et al., 2002) appeared that overcame the limitations of the other Cognitive Diagnostic

models. The next sections describe each of these Cognitive Diagnostic models in detail.

2.2.1 Fischer's LLTM

Linear Logistic Trait Model (LLTM) is one of the oldest cognitive models, proposed in 1973 by Fisher, who decomposed item difficulty parameters of the logistic model into discrete cognitive-attribute-based difficulties. The Rasch item difficulty, also called *effect*, equals "the weighted sum of these attribute-based difficulties," as made known in Equation (7) (Fischer, 1973):

$$\sigma_i = \sum_k f_{ik}\eta_k + c \tag{7}$$

where the weight of factor k in item i, denoted as f_{ik}, indicates whether factor k is required by item i; η_k is the effect or difficulty parameter of factor k across the entire exam; and c is an unknown constant (Fischer, 1973). The Rasch item difficuly or *effect*, denoted as $\sigma_{I,}$, is the difficulty parameter of factor k across the whole exam(Fischer, 1973). By replacing this Rasch item difficulty (σ_i) for the item difficulty parameter in the logistic model, the LLTM was developed as shown in Equation (8):

$$P\left(X_{ij}=1\middle|\theta_j\right)=\frac{1}{1+e^{-\left(\theta-\left(\sum_k f_{ik}\eta_k+c\right)\right)}} \tag{8}$$

where X_{ij} equals 1 when examinee j answers item i correctly and equals 0 when examinee j answers item i incorrectly.

These discrete cognitive-attribute-based item difficulties were used as one kind of parameter showing the basic cognitive operations for correctly solving an item (Fischer, 1973). These operations can be considered as attributes or cognitive building blocks. Therefore, LLTM is considered to be an item-based cognitive model.

Even though η_k is on the θ_j scale and examinees can be cognitively diagnosed via unidimensional proficiency scaling using LLTM, the model lacks some important properties. First, a unidimensional ability parameter remains as a single unidimensional parameter (θ_j) and lacks the measure for evaluation of individual examinees with respect to the individual attributes (Hartz, 2002). Second, the difficulty parameter η_k does not indicate the difficulty of each attribute for each suitable item (Hartz, 2002). This parameter only indicates the difficulty of an attribute across the entire exam.

2.2.2 Rule Space Methodology

Unlike the LLTM, the Rule Space Methodology, developed by Kikumi Tatsuoka and her associates (K.K. Tatsuoka, 1983, 1984, 1990, 1995; K.K. Tatsuoka & M.M. Tatsuoka, 1982, 1984; M.M. Tatsuoka & K.K. Tatsuoka, 1989), generates diagnostic scores. This approach can be explained as a decomposition of examinee abilities into cognitive components (Tatsuoka & Tatsuoka, 1982). These diagnostic scores can be characterized by defining a vector of attributes α. An attribute in a measurement can be categorized as a skill, knowledge, task, subtask, or cognitive process that an examinee may or may not have (Tatsuoka, 1995). The Rule Space Methodology overcomes the difficulties in Cognitive Diagnosis arising from the unobservable characteristics of cognitive processes and knowledge states. Therefore, the relationship between the items on a test and the attributes that they measure needs to be determined.

The Rule Space Methodology is composed of two parts: Q-matrix theory and rule space. Q-matrix theory determines unobservable knowledge states and changes them into observable item response patterns (Tatsuoka, 1995). This operation involves establishing the relationship between the items and the attributes they are measuring. These attributes may or may not be grasped by the examinee. This mastery or non-mastery of the attributes by the individual examinee is represented in an attribute vector called the knowledge state (Tatsuoka, 1990; Tatsuoka, 1995).

In the second part of the methodology, a classification space, called the rule space, is constructed to classify an examinee's item responses into one of the

knowledge states that are established in the first part (Birenbaum & Tatsuoka, 1993; Tatsuoka, 1995). A rule space represents a set of the examinee's item responses. These item responses can be related to the specific knowledge state. Using vectors of attribute mastery/non-mastery (dichotomously coded), the procedure classifies the examinees (Tatsuoka, 1983, 1995). Moreover, these vectors are hypothesized to generate the observed item responses stochastically (Tatsuoka, 1983, 1995). It is assumed that cognitive "rules" govern the observed "stochastic" response patterns (Tatsuoka & Tatsuoka, 1982). These rules, later identified as a Q-matrix, establish an "ideal response pattern" (Tatsuoka, 1990, 1995). An ideal response pattern is one that is obtained through a particular hypothetical combination of mastery and non-mastery levels of the attributes (Tatsuoka, 1995). The comparison between the Q-matrix and students' observed item response patterns is completed by establishing a series of ideal response patterns. This means that, when an examinee has mastered all required attributes for a certain item, the examinee is supposed to answer that item correctly. When an examinee is deficient in at least one of the required attributes, however, the examinee is supposed to answer that item incorrectly (Tatsuoka, 1990, 1995).

Equation (9) shows the deterministic model of the ideal response pattern:

$$P(X_{ij} = 1 \mid \text{examinee } j) = \begin{cases} 1 & \text{if } \underline{\alpha} = \underline{1} \\ 0 & \text{if } \underline{\alpha} \neq \underline{1} \end{cases} \qquad (9)$$

The observed response pattern (which is generated stochastically) and the ideal response pattern (which is based on cognitive "rules") are then compared to see whether differences are found. The distance between these two patterns is used for the classification of the examinee (Birenbaum & Tatsuoka, 1993; Hartz, 2002). So-called attribute mastery patterns are represented as a vector of ones and zeros.

Q-Matrix

One of the significant developments from Tatsuoka's Rule Space approach is the Q-matrix (K.K. Tatsuoka, 1990), which consists of coding for which items measure which attributes that are necessary to solve the problem. The Q-matrix is also referred to as an "incidence matrix" in which rows represent attributes and columns represent items. Let's say that the Q-matrix is a K x n matrix containing ones and zeros, where K indicates the number of attributes we wish to assess and n indicates the number of items on the test. Each cell of the Q-matrix, q_{ik}, takes a value of 1 if mastering skill k is required to solve item i, and 0 otherwise.

Below is an example of a 4 x 5 Q-matrix:

	*item*1	*item*2	*item*3	*item*4	*item*5
*attribute*1	1	0	1	1	1
Q = *attribute*2	0	0	1	1	1
*attribute*3	1	0	1	0	0
*attribute*4	1	1	1	0	0

According to this Q-matrix, the attributes 1, 3, and 4 must be mastered by examinees to solve item 1. That is, the first item measures attribute 1, attribute 3, and attribute 4, all of which must be mastered by each examinee. The second item measures attribute 4 only, while the third item measures all four attributes. Each of the fourth and fifth items requires attributes 1 and 2. This Q-matrix is used widely in Cognitive Diagnosis Theory because it has the advantage of capturing the relationship between items and attributes.

Even though Rule Space Methodology is one of the fundamental Cognitive Diagnosis theories, it has not been widely accepted for several reasons. In the practical world, students do not behave exactly according to the theory reflected in the Q-matrix. Also, the theory does not provide any evaluation of the relationship between the items and the attributes (Hartz, 2002). The Q-matrix is usually constructed by content experts, teachers, and researchers in the related field for more accurate analysis. Without the evaluation, however, it would be difficult to investigate whether a user-specified Q-matrix is sufficiently representing the attributes required by each item (Hartz, 2002). Therefore, many other models have been proposed to explain the uncertainty in this methodology.

2.2.3 The Unified Model

The Unified Model, developed by DiBello, Stout, and Roussos (1995), is based on the Rule Space Methodology of Tatsuoka & Tatsuoka (1982), which

decomposes examinee abilities into cognitive components, as well as on Fischer's LLTM (1973), which decomposes the item difficulty parameter into discrete attribute-based difficulties. As a result, the Unified Model features both item-based attribute parameters and examinee-based attribute parameters. This model simultaneously combines the discrete, deterministic aspects of cognition that lie beneath Cognitive Diagnosis theory and the continuous, stochastic aspects of test response behavior that characterize IRT (DiBello et al., 1995).

The Unified Model is illustrated in Equation (10):

$$P\left(X_i = 1 \middle| \underline{\alpha}_j, \theta_j\right) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \cdot q_{ik}} r_{ik}^{(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\theta_j) + (1 - d_i) P_{b_i}(\theta_j) \qquad (10)$$

where α_{jk} denotes examinee j's mastery of attribute k, where a 1 represents mastery and a 0 represents non-mastery. The factor q_{ik} is the Q-matrix entry for item i and attribute k, and θ_j is the latent residual ability (DiBello et al., 1995). The probability of $P(\theta_j)$ derives from the Rasch model, with the item difficulty parameter denoted by the subscript of P. The parameter d_i is the probability that the Q-based strategy is selected over other possible strategies (DiBello et al., 1995).

The parameter π_{ik} is the probability that examinee j will correctly apply attribute k to item i given that examinee j does possess attribute k (DiBello et al.,

1995). Mathematically, this is written as Equation (11), with Y_{ijk} equaling unity when the correct application of the attribute is present.

$$\pi_{ik} = P(Y_{ijk} = 1 | \alpha_{jk} = 1) \quad (11)$$

Last, the parameter r_{ik} is the probability that examinee *j* will correctly apply attribute *k* to item *i* given that examinee *j* does not possess attribute *k* (DiBello et al., 1995).

$$r_{ik} = P(Y_{ijk} = 1 | \alpha_{jk} = 0) \quad (12)$$

A fundamental difference between Tatsuoka's (1983, 1993) Rule Space approach and the Unified Model lies in their attempts to model in some detail. The Unified Model contains the sources of systematic deviations from the response behavior predicted by the Q-matrix (DiBello et al., 1995). In Tatsuoka's model, however, the deviations from responses predicted by the Q-matrix are modeled in terms of the standard item response probabilities based on the usual IRT latent ability (DiBello et al., 1995).

In Tatsuoka's model, the source of random errors is considered random slips, with the result that all systematic errors are considered to be random slips. In the Unified Model, however, systematic error is broken down into four types (DiBello et al., 1995).

(1) Strategy selection: The response variation derives from the selection of strategy. When an examinee answers an item, he/she can use a different strategy than the one captured in the Q-matrix.

(2) Completeness of the Q-matrix: If an item measures an attribute that is not specified in the Q-matrix, then the Q-matrix would be considered incomplete.

(3) Positivity: *Positivity* can be defined as the inconsistency of student responses. Two examples of inconsistency are the case in which students possessing a certain attribute do not apply it correctly and answer incorrectly an item measuring that possessed attribute, and the case in which students not possessing a certain attribute answer correctly the item that measures the possessed attribute. The value of positivity is high for individuals who possess an attribute and use it correctly and for individuals who do not possess an attribute and fail to use it correctly.

(4) Slips: Slips is a category of random error that cannot be explained by the above three categories. This category includes mental glitches resulting in careless mistakes (bubbling in a wrong multiple-choice option, forgetting to put positive/negative signs, and so on), even though the student correctly solved the problem.

There are important advantages in using the Unified Model, one of which is the use of an important parameter, θ_j. The Unified Model θ_j is totally different from the IRT-based θ, which represents the examinees' individual abilities as a whole. Samejima (1995) thought of this θ_j as the parameter that can measure higher mental processes. DiBello et al. (1995) thought of it as a nuisance parameter that can deal with multiple solution strategies, or strategies that cannot be measured by the Q-matrix. This parameter measures the ability construct that is left over, that is, the construct not included in the Q-matrix. Let's say that the latent ability space consists of $\underline{\alpha}_Q$ and $\underline{\alpha}_b$ (Hartz, 2002). This $\underline{\alpha}_Q$ is the latent ability explained by the Q-matrix, and $\underline{\alpha}_b$ is the remaining latent ability that cannot be explained by $\underline{\alpha}_Q$. The important parameter θ_j of the Unified Model is set to measure the remaining latent ability $\underline{\alpha}_b$, while the parameter $\underline{\alpha}_j$ is set to measure $\underline{\alpha}_Q$.

One might say that this approach simply adds more attributes into the Q-matrix; however, more attributes cannot be added just to account for the residual abilities, because adding more parameters into the model would greatly complicate the measurement process (McGlohen, 2004). In addition, the more attribute parameters to be estimated, the more items are needed to get acceptable reliability (DiBello et al., 1995). For the sake of parsimony of test time and test length, DiBello et al. (1995) developed the new latent residual ability parameter θ_j to capture the abilities not measure by Q-matrix.

Another advantage of the Unified Model is its unification of the deterministic approach and the stochastic approach. For each examinee, the model includes both a latent ability θ and a latent attribute state α. In addition, it includes a deterministic approach to estimating knowledge state to assess examinees with respect to the underlying attributes, and it includes a stochastic approach to examining the relationship between the items and the attributes (DiBello et al., 1995). It is said that the deviation probabilities from Q-predicted responses clearly depend on both θ and the examinee's cognitive state α (DiBello et al., 1995). Thus, each item response function is defined directly in terms of θ and α.

Practically, however, the Unified Model contains some parameters that are not identifiable. The item parameters of the Unified Model need to be estimated for the model to be calibrated. It is necessary, therefore, to reduce the parameter space before estimating its parameters (Hartz et al., 2002). Currently, the Unified Model lays a solid foundation for a Fusion Model, because it provides flexibility and interpretability of parameters.

2.2.4 The Fusion Model

Among the models of the Cognitive Diagnosis approach, the Fusion Model, developed by Hartz, Roussos, and Stout (2002), is considered highly successful because it satisfies three conditions necessary for a model to be effective. These conditions, identified by Hartz et al. (2002), are that the model:

(1) Give an estimation of examinee attributes

(2) Relate items to attributes

(3) Provide statistical identification of the model's parameters

The foundation of the Fusion Model is the Unified Model, which features flexibility and interpretability of item parameters. The Unified Model, however, does not have statistically estimable parameters (Hartz, 2002). Therefore, the Fusion Model is used for the research described in this dissertation. This model is considered to be a statistically tractable item response model with parameters that represent both the cognitive profiles of the examinee and the item relationships to these cognitive attributes (Hartz et. al, 2002). The advantage of the Fusion Model is statistical identifiability by reducing the number of parameters involved in the modeling (Hartz et. al, 2002).

To repeat, the IRT-based Cognitive Diagnosis models define the probability of observing the response of examinee *j* to item *i* given the examinee's ability parameters and item parameters (Hartz, 2002). This probability is denoted as $P\left(X_{ij} = x|\underline{\theta}_j \underline{\beta}_i\right)$. The symbol $X_{ij} = x$ indicates the response of examinee *j* to item *i*, where $x = 1$ indicates a correct response and $x = 0$ indicates an incorrect response. The term $\underline{\theta}_j$ indicates a vector of examinee *j*'s ability parameters, and $\underline{\beta}_i$ indicates a vector of the item parameters (Hartz et. al, 2002).

The characteristic that distinguishes the new versions of cognitive diagnosis models, such as the Unified Model and the Fusion Model, from other IRT models is that the items (i = 1,…, I) are related to a set of cognitive attributes (k = 1,…, K) (Hartz et al., 2002). The Fusion Model is an attempt to evaluate each examinee with respect to each attribute. These relations can be observed as f_{ik}, the weight of attribute k in item i (Fischer, 1973). This also can be simplified to the Q-matrix first introduced by Tatsuoka (1990). As already explained, the Q-matrix can be described as $Q = \{q_{ik}\}$, where q_{ik} = 1 indicates that attribute k is required by item i, and $q_{ik} = k$ indicates that the attribute k is not required by item i (Tatsuoka, 1990, 1995).

Equation (13) shows the resulting Fusion Model item response function, which is based on the same examinee parameters—$\underline{\alpha}_j$ and θ_j—as those from the original Unified Model.

$$P\left(X_{ij} = 1 \middle| \underline{\alpha}_j, \eta_j\right) = \pi^*_i \prod_{k=1}^{K} \left(r^*_{ik}{}^{(1-\alpha_{jk})}\right)^{q_{ik}} P_{c_i}(\eta_j) \qquad (13)$$

Fusion Model Item Parameters

The parameter π_i^* is the probability of correctly applying all item *i* required attributes, given $\alpha_{jk} = 1$, which means that examinee *j* has mastered attribute *k.* In other words, the probability pertains to whether an examinee who has mastered all attributes for item *i* can correctly apply those attributes when solving for item *i*. The term can be interpreted as the Q-based item *i* difficulty (Hartz et al., 2002).

$$\pi_i^* = \prod_{k=1}^{M} \pi_{ik}^{q_{ik}} \tag{14}$$

The parameter r_{ik}^* is the proportional parameter representing the ratio of the likelihood of a correct answer given mastery versus non-mastery (Hartz et.al, 2002). In Equation 15, Y_{ijk} = 1 indicates that attribute *k* is correctly applied to examinee *j* for item *i*, and 0 otherwise. The equation α_{jk} = 1 specifies that examinee *j* has mastered attribute *k*, and 0 otherwise.

$$\begin{aligned} r_{ik}^* &= \frac{P\left(Y_{ijk} = 1 \mid \alpha_{jk} = 0\right)}{P\left(Y_{ijk} = 1 \mid \alpha_{jk} = 1\right)} \\ &= \frac{r_{ik}}{\pi_{ik}} \end{aligned} \tag{15}$$

The numerator is really the probability that examinee *j* correctly answers item *i* given examinee *j* has not mastered attribute *k*. The denominator is the probability that examinee *j* correctly applies attribute k to item i given examinee *j* has mastered attribute *k*. Therefore, r_{ik}^{*} is interpreted as the item *i* discrimination parameter for attribute *k*. It represents the penalty for lacking attribute *k*, a comparison between the correct item response probabilities of lacking attribute *k* and of mastering attribute *k* (Hartz et.al, 2002). A high r_{ik}^{*} value signifies that attribute *k* is not important in producing a correct response to item *i* (Hartz ,2002; Hartz, et al., 2002). Therefore, the closer r_{ik}^{*} is to zero, the more discriminating item *i* is for attribute *k*.

In equation (13), the term $P_{c_i}\left(\eta_j\right)$ represents the fact that the Q-matrix does not include all the related attributes (Hartz et. al, 2002). The parameter c_i is equivalent to the amount that the correct item performance requires, η_j, in addition to the required Q attributes; this c_i parameter refers to the completeness index for item *i* (Hartz et. al, 2002). This is a most unique and important component preserved from the Unified Model due to the fact that Q-matrix cannot include all relevant cognitive attributes (Hartz et. al, 2002).

Fusion Model Person Parameters

According to Hartz (2002), mastery of one attribute is statistically dependent on mastery of another attribute. Also, an examinee who has mastered many

attributes among $\underline{\alpha}$ is expected to have a higher θ. Abilities for examinee *j* is projected as ($\underline{\alpha}_j$, θ_j). The parameter $\underline{\alpha}_j$ is a vector of 0's and 1's representing whether examinee *j* has mastered attribute *k* ($\alpha_{jk} = 1$) or has not mastered attribute *k* ($\alpha_{jk} = 0$). The parameter θ_j is a continuous variable indicating a unidimensional projection of examinee *j*'s residual ability that cannot be measured by Q-matrix (Hartz et. al, 2002).

The Fusion Model incorporates $\tilde{\alpha}_{jk}$ when *k* = 1,…, *K*, with standard normal priors distribution that generates dichotomous attributes rather than dichotomous items (Hartz, 2002). That means normally distributed variables generate mastery versus non-mastery dichotomous attribute (Hartz, 2002). The vector of $\tilde{\alpha}_{jk}$ needs to be converted to the dichotomous α_{jk} using k_k, the "cutoff" for mastery of attribute *k*. Note that $\alpha_{jk} = 1$ when $\tilde{\alpha}_{jk} > k_k$, which means that the examinee has mastered the attribute *k* because the examinee's latent ability is greater than the cutoff value for mastery (Hartz, 2002). Equation (16) shows the representation of the converted attributes.

$$\alpha_{jk} = \begin{cases} 1 & \tilde{\alpha}_{jk} > k_k \\ 0 & \tilde{\alpha}_{jk} < k_k \end{cases} \qquad k = 1,\ldots, K \qquad (16)$$

So far, a very fundamental test theory, IRT, and a newly developed test theory, Cognitive Diagnosis, have been discussed in detail. The characteristics of IRT, especially test information, and the new parameters of the Cognitive Diagnosis models described above are used as important features of the automated test assembly method that combines both fields of psychometrics. The next section gives a general discussion of automated test assembly and develops the specific algorithm used in the research for this dissertation.

2.3 AUTOMATED TEST ASSEMBLY METHODS

2.3.1 Traditional Test Construction

The IRT approach to test development concentrates on the item and test information functions. Item Information Functions (IIFs) are the building blocks of IRT test assembly. Lord (1977, 1980) and Birnbaum (1968) suggested that the properties of IIFs were useful because these curves could be added together to approximate the desired shape of the TIF. In other words, tests can be assembled to fit the specified shape of the TIF by using the additive properties of the IIF curves. Items are selected based on the amount of information each item contributes (item information function) to the amount of information of the test as a whole (target test information function). Once the item parameters have been estimated, item selection is very straightforward.

Lord (1980) summarized the following procedure for constructing tests under the IRT framework.

1. A target test information function (TTIF) is set up by specifying that the TIF be a certain shape based on the purpose of the test. TTIF is the test information function for an optimal set of test items.
2. Test items with item information functions that meet the requirements of the target function are selected for the TIF.

3. The TIF is computed as the item information function curves of the items are added to the test. This step determines the information contribution of the items to the TIF.
4. Items are continually added to the test so that the computed TIF approximates, in some acceptable sense, the target TIF for the specified ability level.

Even though this method is widely used, it has some limitations. One major limitation is that it is hard to achieve a desired TIF when the item bank is large (Thuenissen, 1985). In Lord's (1980) description of test generation, it is assumed that item selection is done by hand. When there are many items in the item bank, however, it is impractical to manually select items for their IIFs to fit a TIF, because tests are usually constructed to meet detailed specifications (Fletcher, 2003). It is difficult to tell whether the optimal set of items that meets all of the specifications has been selected. In addition, as Fletcher (2003) observes, when one assembles tests under the IRT framework, the task of taking into consideration all the test requirements and specifications is laborious and computationally inflexible. There is a solution to this limitation, however: the use of automated test assembly methods.

2.3.2 Automated Test Assembly

The widespread use of computers in educational and psychological measurement, as well as efforts to satisfy test specifications, has led to the need for automated test assembly, a method that can be useful for large-scale testing. In that case, it is necessary to compare the two requirements just mentioned with those for assembling linear test forms in large-scale testing. Tests can be constructed automatically by the application of mathematical programming models (Adema, 1992; Adema & van der Linden, 1989; Baker, Cohen, & Barmish, 1988; Boekkooi-Timminga, 1987; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1987). These mathematical programming Methods need to have an optimal solution, a solution that satisfies both the conditions of the problem and the given objective (Gass, 1985).

In such tests, rather than assembling one individual test form at a time, it is common to assemble a set of parallel test forms from an item pool at the beginning of a new planning period (van der Linden, 1998). Because large-scale assessments are intended for large groups, it is important to have a set of parallel test forms to give out at all testing sessions and locations. If the assessment is based on IRT, the forms can be defined as parallel insofar as each of them contains (1) the combinations of items needed to meet all test specifications for the test, and (2) the items that are most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for the population of examinees (van der Linden, 1998).

As mentioned above, test specifications are used as the foundation for test assembly. A distinction is made between two types of test specifications, objectives and constraints. Objectives require that a test attribute or a function of item attributes take a minimum or maximum value (van der Linden, 1988, 1994). For that reason, objectives can be formulated as mathematical functions to be optimized (van der Linden, 1998). The term that is optimized is defined as the objective function of the optimization problem. It can be said that a linear-programming problem has a linear function of the variables to aid in choosing a solution to the problem (Gass, 1975). This linear combination or mathematical function of the variables, called the objective function, must be optimized by the selected solution. For a given item bank, an objective involves its own optimal combination of items; in other words, some aspect of the items to be selected for a test is optimized. Thus, test assembly programs can optimize only one objective function at a time.

Constraints require that a test attribute or a function of item attributes meets an upper and/or lower limit (van der Linden, 1998). These constraints can be formulated as mathematical equalities or inequalities. In contrast to objectives, the number of constraints is unlimited, depending on the test specifications and test conditions. As regards the constraints, test constructors have a variety from which to choose. The three main types of constraints are those that deal with categorical item characteristics, those that deal with quantitative features of the items, and those that deal with inter-item dependencies (van der Linden, 1998; van der Linden, 2000).

With the linear program, the test assembly problem can be formulated as an optimization problem. In that case, it is necessary to specify the test assembly as an example of constrained combinatorial optimization (van der Linden, 1998). A test assembly problem, therefore, can be described as a combination of an objective with a set of constraints. As modified from its use in operational research, the linear program Method minimizes or maximizes an objective function while satisfying a series of linear constraints.

The primary goal of the linear program Method is to maximize or minimize an objective function (for example, to minimize the number of items in the test, to maximize information at a certain point on the ability scale, or to maximize the reliability of the test) so that test specifications (for example, test length, test information, content area, and item format) are met in the form of optimal constraints (for example, that the length of the test equals 30 items, that test information has upper and lower bounds, or that the mean *p*-value equals 0.25) (Fletcher, 2003).

2.4 ZERO-ONE (0/1) BINARY LINEAR PROGRAMMING METHODS

One method for taking advantage of the psychometric properties of IRT while meeting complex test specifications is mathematical linear programming (LP) in the form of Zero-One Binary (0/1) Programming (see, for example, Adema & van der Linden, 1989; Boekkooi-Timminga, 1987, 1990; de Gruijter, 1990; Theunissen, 1985, 1986; van der Linden & Boekkoi-Timminga, 1988, 1989). Because test assembly involves either the inclusion or exclusion of a specific item, 0/1 Linear Programming

is one way of approaching automated test assembly. For the further studies of 0/1 Linear Programming, please refer to the textbook of Gass (1985), Jensen and Bard (2003), or Bertsimas and Tsitsiklis (1997).

Tests are seldom assembled by only matching a target information function in the manner described for Birnbaum's method (1968). The tests, moreover, are assembled with attention to a numbers of test specifications, including content balancing, item format, section length, test length, reliabilities, word counts, and many more (van der Linden, 1998). Among many researchers working with 0/1 Linear Programming, Theunissen (1985) was the first to present a 0/1 Linear Programming Method for the construction of a test of a target information function with minimization of the test length. Therefore, a test of minimal length was constructed using a branch-and-bound algorithm. In his method, the objective function shows the minimization of test length, and the test information function lies above the target function at a number of ability points chosen in advance. Several other studies explored this issue, for example, Boekkooi-Timminga (1987); Boekkooi-Timminga and van der Linden (1987); Theunissen (1986); and van der Linden and Boekkooi-Timminga (1988).

The advantages of 0/1 Linear Programming is its flexibility and simple modeling steps (Gass, 1985; Jenson & Bard, 2003). Most of the problems can be modeled using 0-1 integer variables. Once a model has been formulated, some computer algorithms or commercial computer programs such as GAMS, CPLEX

(ILOG, 2003) can be used to solve the linear programming problems. The solution can be found using the branch-and-bound method.

The branch-and-bound method is the basic workhorse technique for solving integer and discrete programming problems (Jawloms. 1988, Nemhauser & Wolsey, 1988). It can be defined as the method to get the optimal solution by keeping the best solution found so far (Jenson & Bard, 2003). If a partial solution cannot improve on the best, it is abandoned. This method can be most easily understood in graphical form.

Figure 1 shows the complete enumeration of all of the solutions. It is a simple branch-and-bound method of selecting two items from three possible items. Three items are shown below as in parentheses: (item1, item2, item3). A one indicates that the item has been selected and a zero indicates that the item has not been selected. A number sign (#) indicates that the decision has not been made yet.

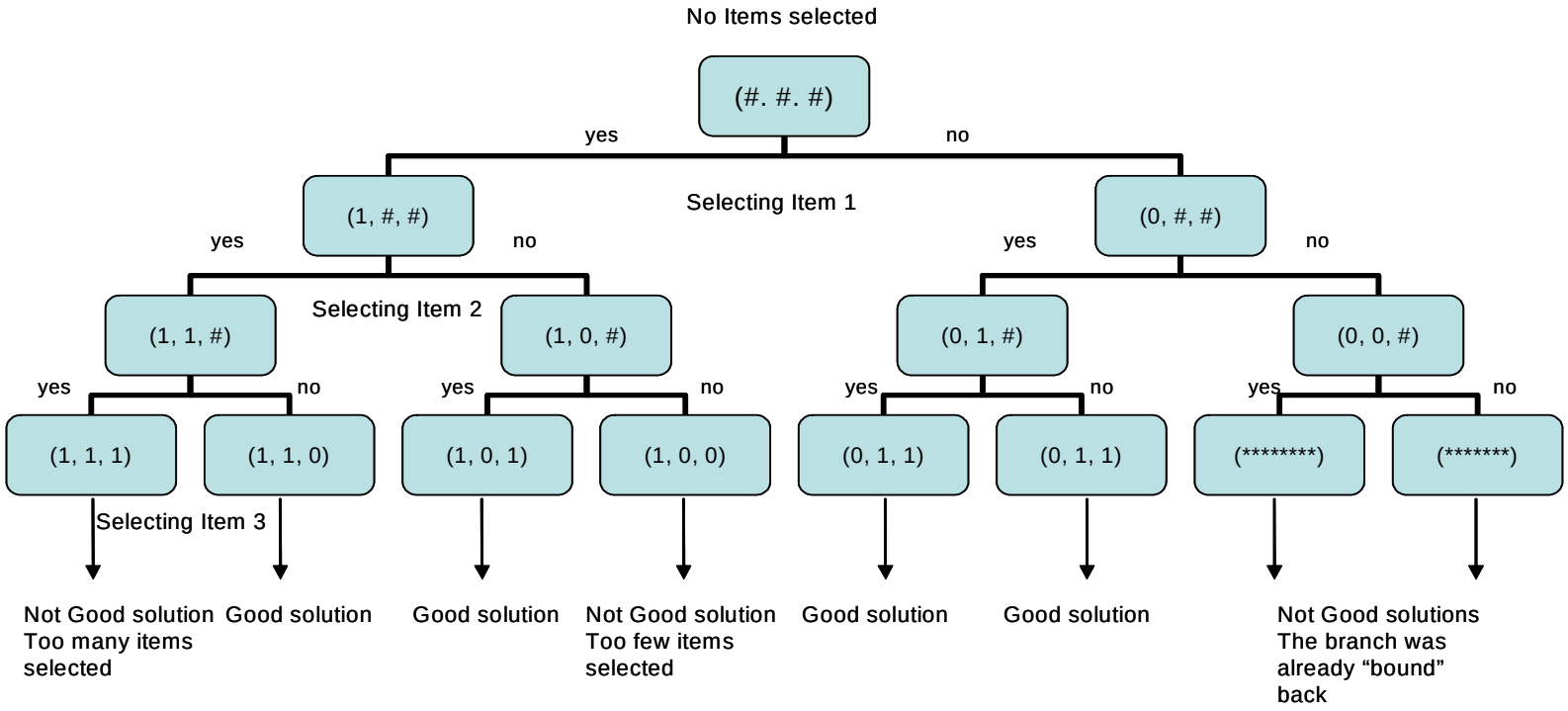


Figure 1: Branch-and-Bound Method of selecting two items.

The branching continues until all the possible items are considered. The first box shows three #s, meaning no items are selected. The first decision to make is either to include or not to include the first item, with two possible branches. Each branch splits into two branches, to decide whether or not to select a certain item. The branching continues until all possible items in the item bank are considered. In this example, eight possible branches result. Among those, any particular branch containing a set of items that does not obey the specified constraints is not investigated any further. This branch is now *bounded*, and it is not permitted to grow any more. After the determination of all possible branches that correspond with all of the requirements in the list of constraints, the objective function is used to select the best combination of items (Hawkins, 1988; Jensen & Bard, 2003).

The following shows an example of a binary programming model that maximizes the information on specific θ values (equations (17) to (21)):

Maximize

$$\sum_{i=1}^{I} I(\theta_i) x_i \qquad (17)$$

Subject to

$$\sum_{i=1}^{N} x_i = n, \qquad (18)$$

$$\sum_{i=1}^{N} a_{ij} x_i \geq L_j \qquad j = 1, \ldots, J, \qquad (19)$$

$$\sum_{i=1}^{N} a_{ij} x_i \leq U_j \qquad j = 1, \ldots, J, \qquad (20)$$

and

$$x_i \in \{0,1\}. \qquad i = 1, \ldots N \qquad (21)$$

Let i = 1, …, I be the index of items in the pool and j = 1, …, J be the index that denotes item properties related with the nonpsychometric constraints. The variable x_i denotes the decision variable that determines whether item i is included in (x_i =1) or excluded from (x_i=0) the test. The terms Lj and Uj are lower and upper bounds on the number of items in the test having each property, respectively. When item i has property j, a_{ij} equals 1 (a_{ij} =1), and if it does not have property j, a_{ij} equals 0 (a_{ij} =0).

In this approach, first an objective function can be specified to either maximize or minimize a certain function. In that case, the information function is maximized according to certain constraints (equations (18) to (21)). Some examples of objective functions include minimizing the largest positive deviation from the target, minimizing the sum of the positive deviation from the target test information, minimizing test length, or minimizing other characteristics of the test (Swanson & Stocking, 1993). Equations (19) and (20) show the nonpsychometric constraints as lower and upper bounds on the number of items in the test with the specified properties. Therefore, in this case, the information is calculated for every combination of items that obeys all of the constraints,

and the combination with the greatest information is selected as the best combination of items.

Other possible forms of the objective functions are explained in van der Linden and Boekkooi-Timminga (1989). Also, refer to van der Linden (1944) or van der Linden (1998) for the further examples. For this dissertation, three ways of optimizing the objective functions are discussed; Minimax Method, Maximin Method, and Maximum Information Method. Before discussing these methods, target information must be addressed.

2.4.1 Absolute Target vs. Relative Target

When a test is constructed using automated test assembly methods, a target for a TIF makes goal values (also called target values) available along the θ scale to use for the item pool. For all kinds of IRT models, the TIFs are smooth, well-behaved functions (van der Linden, in press). Therefore, if a TIF is required to meet a smooth target function at one θ point, it can be said that the TIF approximates the target at the θ points that are next to an actual pick-up θ point. van der Linden (in press) suggested having three to five well-chosen points to control the TIF. One of the reasons for choosing a smaller number of points is that fewer points result in much faster solutions in the practical automated test assembly (van der Linden, in press).

In practice, the target values are specified at a few points on the θ scale, also denoted as θ_l, $l = 1,\ldots, L$. These few θ points are assumed to be selected by the test constructors. van der Linden (in press) provided target values that yield excellent results

for the 3PL model: $(\theta_1, \theta_2, \theta_3)$ = (-1.0, 0, 1.0) or $(\theta_1, \theta_2, \theta_3, \theta_4)$ = (-1.5, -.5, .5, 1.5). These points are used as the target values for this dissertation research.

When one uses TIFs as the objective functions, two kinds of targets should be considered: an absolute target and a relative target. The absolute target specifies a fixed number of information units at the θ_l points (van der Linden, 1987). Therefore, more information than the specified target information is not necessary. One example of an absolute target can be found in the long-term testing programs with the goal to assemble test forms that are parallel to a fixed reference form (van der Linden, 1987, in press). T_l is used to denote the absolute target value for the TIF at θ_l, $l = 1,\ldots, L$.

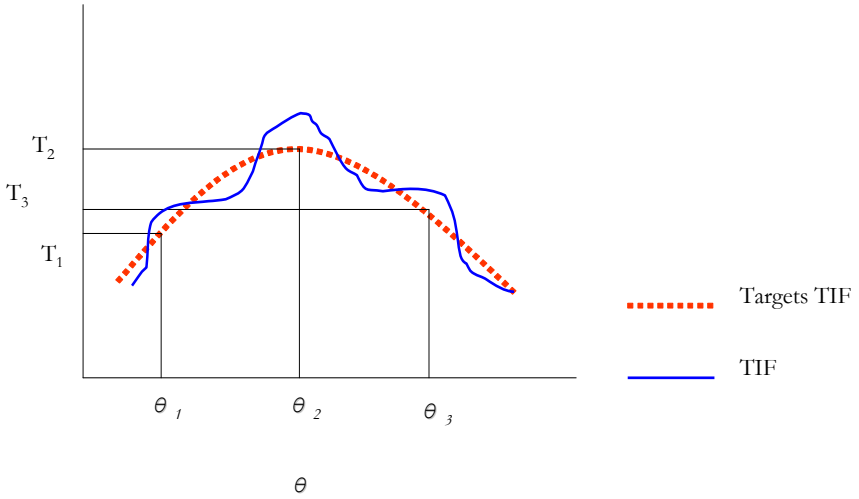


Figure 2: Graphical depiction of an absolute target

A relative target for a TIF concerns the shape of the target but not its height (van der Linden & Boekkooi-Timminga, 1989). Therefore, the more information there is along the θ scale, the better the results of the test construction, as long as the information represents the objectives of the tests. Examples of relative targets include broad-range diagnostic testing and licensure testing with a fixed minimum value for the level of passing performance (van der Linden, in press). The relative target can be symbolized as a set of numbers R_l that represent the essential amount of information at θ_l relative to the other points in the set $l = 1,\ldots, L$ (van der Linden & Boekkooi-Timminga, 1989).

For example, if the test must include three times as much information at θ_l as at θ_{l+1}, then, $\frac{R_l}{R_{l+1}} = 3$. This means that the number R_l should be three times as large as the number R_{l+1}. In contrast to the absolute target, the more important consideration for the relative target is specifying the ratios of the numbers(van der Linden, in press). Therefore, it is not necessary to know the unit of the information measure (van der Linden, in press).

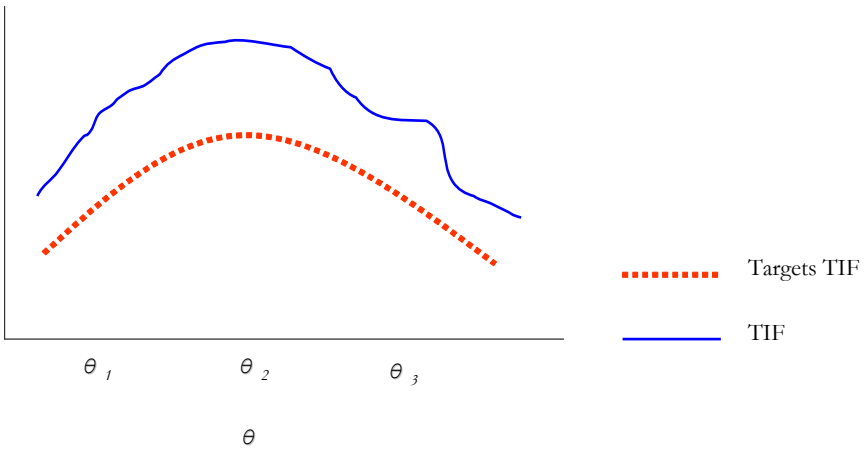


Figure 3: Graphical depiction of a relative target

As currently practiced, the process of test construction employs a combination of IRT, computer processing, and mathematical programming or heuristic methods (see, for example, Adema, 1992; Baker, Cohen, & Barmish, 1988; de Gruijter, 1990; Swanson & Stocking, 1993; Timminga & Adema, 1995; Theunissen, 1985, 1986; van der Linden, 1996; van der Linden & Boekkoi-Timminga, 1989; van der Linden & Luecht, 1996; van der Linden & Reese, 1998). Among examples of the application of test construction processes to a variety of problems are IRT-based test assembly, classical test assembly, multiple test forms assembly, observed-score equating. Generally, these combinations of processes have been used successful, but typically they have not been applied to test construction problems involving Cognitive Diagnostic constraints on item selection. For the research for this dissertation, therefore, an automated test assembly method is implemented to construct Cognitive Diagnosis tests using 0/1 Linear Programming. The next section provides a more detailed description of 0/1 Linear Programming.

2.4.2 Minimax Method: Assembling Tests to Absolute Targets

When absolute targets are used to assemble tests, a fixed number of information units are needed at the θ_l points. This method is essential for the case when the test is assembled in a program where a fixed target has to be preserved over time (van der Linden, 1987). With this constraint, neither positive nor negative deviations from the target values are desirable (van der Linden, in press). van der Linden (in press) recommended the following Minimax Method to assemble the parallel tests by minimizing the largest deviation from the target. The method forces the TIF down as

closely as possible against the target. The following constraints are added to the method to induce the TIF to be similar to the target:

Minimize y (22)

Subject to

$$\sum_{i=1}^{I} I_i(\theta_l)x_i \leq T_l + c_l, \qquad l = 1,\ldots, L \qquad (23)$$

$$\sum_{i=1}^{I} I_i(\theta_l)x_i \geq T_l - d_l, \qquad l = 1,\ldots, L \qquad (24)$$

Let $c_l \geq 0$ and $d_l \geq 0$ be small tolerances within which the TIF is allowed to be larger or smaller than the target values T_l (van der Linden, 1987, in press). The largest absolute deviation from T_l is minimized in Equation (22).

2.4.3 Maximin Method : Assembling Tests to Relative Targets

The Maximin Method was developed by van der Linden and Boekkooi-Timminga (1989). This method involves the selection of items such that they maximize the information from the test, without changing the desired shape of the resulting test information function (van der Linden and Boekkooi-Timminga, 1989; van der Linden, in press). To use this method, the test constructor specifies the relative shape of the target

TIF by selecting R_l. To elicit the relative shape of the target information function from the test constructor, the method requires the following steps (Adema, 1992):

Step 1) The test constructor is made aware of the ability scale by which the items are organized in the item bank. Then, the constructor is given a line displaying the content of the items with location at the same chosen points.

Step 2) The constructor selects the number of scale points he or she wants to consider. There are no restrictions on the number of points and their spacing; θ_l, l = 1,…, L, denotes each point.

Step 3) A fixed number of chips (for example, 100) is given to the test constructor. The constructor is then asked to distribute these chips over the scale points so that they reflect the relative distribution of the information that the target test is intended to assess.

Step 4) The test constructor is asked how many items he or she desires in the test.

In van der Linden and Boekkooi-Timminga's article (1989), the variable R_l is said to be the number of chips (the amount of information) the test constructor puts at a certain point θ_l (l = 1,…, L). The relative target information function is characterized by a series of lower bounds (R_1y, …, R_ly). Here, a dummy variable y is maximized by becoming an "explicit common lower bound" to the relative information $R_l^{-1}\sum_{i=1}^{I} I_i(\theta_1)x_i$ at θ_l points (van der Linden, in press).

Maximize y (25)

Subject to

$$\sum_{i=1}^{I} I_i(\theta_l)x_i - R_l y \geq 0, \qquad l = 1,\ldots, L \tag{26}$$

$$y \geq 0. \tag{27}$$

2.4.4 Maximum Information Method: Cutoff Scores

For tests that are used for decision making with a cutoff score, θ_c, only "informative estimates $\hat{\theta}$" are needed in the vicinity of θ_c (van der Linden, in press). The objective function is optimized at the θ_c point, which means that the relative target for TIF narrows to a simple maximization at θ_c (van der Linden, in press). The information at all other points of θ is ignored. The objective function is described in equation (28), which replaces θ with θ_c.

Maximize

$$\sum_{i=1}^{I} I_i(\theta_c)x_i \tag{28}$$

In this study, the three methods described above—Minimax, Maximin, and Maximum Information—are used to select the best items for the traditional IRT tests when they are combined with aspects of Cognitive Diagnostic modeling. Practical constraints on this approach, as well as the actual methods, are explained later. The following chapter discusses the procedure and the data simulation of this study.

The overall purpose of this research is to develop an algorithm for generating information-rich tests by combining Cognitive Diagnosis models with the traditional IRT approach. Such a combination not only produces a single score to measure an examinee's ability level but also provides diagnostic information. Thus, this study describes a new method of automated test assembly, one that incorporates diagnostic techniques with existing IRT-based testing assembly methods. Because of the NCLB Act (2001), educators, parents, and students are increasingly seeking helpful and constructive feedback to learners in educational assessments wherever measures are made of content domains. Therefore, it has become essential to acquire good test development methods, and one of the most attractive of these methods may involve Cognitive Diagnosis modeling.

CHAPTER 3: METHODOLOGY

Traditional methods of automated test assembly involve a variety of test specifications, including content balancing, item format, section length, test length, reliabilities, count of words, and many more (van der Linden, 1998, 2000, in press). The combinations of these constraints have been successful so far; however, the test construction problems involving Cognitively Diagnostic constraints on item selection have only recently been addressed. In this dissertation, an automated test assembly method is implemented that constructs Cognitively Diagnostic tests using 0/1 Linear Programming.

The purpose of this dissertation is to combine the newly developed approaches of diagnostic assessment with the existing approaches of IRT assessment and thereby provide examinees not only their already widely accepted ability-level scores but also helpful *diagnostic* information. Other approaches to automated test assembly provide one or the other means of assessment for constructing tests, but not both. Therefore, a major goal of this research is to identify a set of the most reasonable constraints in Cognitive Diagnosis and to integrate those new constraints into traditional IRT scaling.

The description of this research first involves the procedure for obtaining item parameters and an item pool. The item parameters are pre-calibrated based on a large-scale assessment. Also in this dissertation a dataset is generated based on each simulee's true ability level and true knowledge state. The data set and parameters are used in combination to determine the response patterns of the simulees to the test.

This chapter describes the new approaches to assembling tests automatically. Three methods for optimizing objective functions are examined: a Minimax Method, a Maximin Method, and a Maximum Information Method. In addition, data generation and the simulation of the testing process using these new approaches are illustrated to evaluate how well the methods work.

3.1 SIMULATION STUDY

3.1.1 Item Pool Structure

Item parameters

The data used for the simulation study were real student responses from the Grade 3 math exams administered by the Texas Education Agency (TEA) in the springs of 2000, 2001, and 2002. Psychometric properties such as reliability and validity were carefully scrutinized by TEA to ensure the quality of the test as well as to create a sound assessment. The reliability coefficients using the Kuder-Richardson-20 ranged between low of 0.8s to high of 0.9s (Texas Education Agency, 2002). Also, TEA inspected the various types of validity, such as content-related validity, construct-related validity and criterion-related validity (Texas Education Agency, 2002). Detailed psychometric data for each year can be obtained from the Texas Student Assessment Program Technical Digest (Texas Education Agency, 2002).

To construct an item bank for the automated test assembly, first, a random sample of two thousand examinees was taken. As mentioned above, three administrations of

math exams were combined to create a bigger item pool. The item response patterns of these examinees for each of the three administrations were used to calibrate the traditional IRT item parameters (*a*, *b*, and *c* parameters). BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to obtain estimates of the item parameters according to the three-parameter logistic model.

The Fusion Model has a software program available, called Arpeggio (Hartz et. al, 2002), which makes the application of the model convenient. To analyze the same response patterns for the Cognitive Diagnosis, the Arpeggio program was used to estimate Fusion Model parameters such as item parameters *p** and *r** and person parameters $\underline{\alpha}_j$ by means of MCMC estimation (Hartz, 2002). The default setting uses four Markov chains, each with a length of four thousand, and one thousand burn-off cycles. Other details of the analysis relating to the default settings are described in the software manual (Hartz et al., 2002).

The mastery/non-mastery status of each of the attributes is also calculated through the Arpeggio run. The attribute mastery parameters of each examinee are estimated as continuous and then dichotomized by evaluating the value in respect to the cut-off value of 0.5 (Hartz et al, 2002). A value greater than or equal to 0.5 is assigned mastery status of the given attribute, whereas, a value less than 0.5 is assigned non-mastery status for the given attribute.

To construct the parallel tests, an item bank is required from which to select items. Each test administration contains 44 math items. To have enough items in the item pool, three administrations of the third-grade TASS mathematics exams were combined for a

total of 132 items (years 2000 to 2002). Then, all the item parameters were included in the pool four times to accommodate an item pool of adequate size, in this case, 528 items.

Q-matrix

The Q matrix, which is based on the test blueprint provided by TEA, was evaluated by McGlohen (2004). There are eleven attributes measured by the math exam, with each item having one attribute. Therefore, the Q-matrix has a single pattern. Appendix B shows the attributes measured by the Q-matrix of items.

The Arpeggio program was used in this analysis as well. Using four thousand burn-off cycles in each of four MCMC chains, the Fusion Model parameters were estimated based on the item response data. The mastery proportions parameters, denoted as p_k, were examined to test whether they had been accurately estimated within a given level of confidence.

Using Arpeggio to calibrate the items, the r^* values were compared. If this value is too high, then the corresponding Q-matrix entry is removed. It is recommended that high r^* values be removed, because a high Q-based item discrimination parameter indicates that the designated attribute is not doing a good job of obtaining a correct response (Hartz, 2002).

3.2 METHODS

This dissertation examined how Cognitive Diagnosis models are used in the traditional IRT approach for assembling tests automatically. Three IRT-based 0/1 Linear Programming Methods address the different objective functions: the Minimax Method, Maximin Method, and the Maximum Information Method. These three methods are described after the next section, which discusses newly developed constraints based on Cognitive Diagnosis. This discussion prepares the way to a better understanding of the methods.

3.2.1 Constraints of Cognitive Diagnosis

Three methods of 0/1 Linear Programming were used in this dissertation research along with the combinations of several constraints required by Cognitive Diagnosis. Some of the practical constraints of Cognitive Diagnosis used in these methods are related to the following: (1) the attributes-related constraints (the number of items on each attribute should be more than 3), and (2) r^* (discrimination).

The first constraint is related to the Q-matrix. Equation 29 shows a constraint of attribute versus item. It gives the lower and upper bounds of the number of items assessing each attribute. Hartz (2002) reports that, according to many researchers, the more attributes used, the more likely they are to yield better student measurements.

A second constraint is related to r^* parameters. Equation 30 shows a constraint of discrimination values. According to Hartz (2002), the parameter r_{ik}^* is considered a

discrimination value of item *i* for attribute *k*. For an examinee lacking a required attribute *k*, her or his correct item response probability is proportional to r_{ik}^{*} (Hartz, 2002), which is identified as the penalty for lacking attribute *k*. The closer r_{ik}^{*} is to zero, the better the discriminating item *i* is for attribute *k*. Therefore, it is better to have a small value to add across the items given to each attribute.

$$LB \leq \sum_{i=1}^{I} q_a x_i \leq UB \qquad a = 1, \ldots, A \qquad (29)$$

$$LB \leq \sum_{i=1}^{N} r_{ik}^{*} x_i \leq UB \qquad k = 1, \ldots, K \qquad (30)$$

After these two constraints were added, three automated test assembly methods were compared to see whether incorporation of the Cognitive Diagnostic elements into IRT-based models was working well. The three automated test assembly methods (Minimax, Maximin, and Maximum Information) are described as following.

3.2.2 Minimax Method

The first automated test assembly method using 0/1 Linear Programming was the Minimax Method:

Minimize y (31)

Subject to

$$\sum_{i=1}^{I} I_i(\theta_l) x_i \leq T_l + y \qquad l = 1,\ldots, L \qquad (32)$$

$$\sum_{i=1}^{I} I_i(\theta_l) x_i \geq T_l - y \qquad l = 1,\ldots, L \qquad (33)$$

$$\sum_{i=1}^{I} x_i = n \qquad i = 1, \ldots, I \qquad (34)$$

$$LB \leq \sum_{i=1}^{I} q_a x_i \leq UB \qquad a = 1, \ldots, A \qquad (35)$$

$$LB \leq \sum_{i=1}^{N} r_{ik}^{*} x_i \leq UB \qquad k = 1,\ldots, K \qquad (36)$$

and

$$x_i \in \{0,1\}, \qquad i = 1, \ldots N \qquad (37)$$

$$y \geq 0 . \qquad (38)$$

Constraints (32) and (33) show the method for the absolute target values. The theta points used in this method are $(\theta_1, \theta_2, \theta_3)$ = (-1.0, 0, 1.0), because van der Linden (in press) suggested that these three values yield results as good as those from calculations with values of four or more. The values were (3, 5, 3) for each of the ability levels.

Constraint (34) is the total number of items in a certain test: *I* is the number of items in the pool, and *n* stands for total test length. For this study, 20 items were selected.

The constraints of Cognitive Diagnosis were added to investigate how well these approaches work with the IRT-based method. Constraint (35) is related to setting lower and upper bounds for the number of attributes for each item. Here, *A* stands for the number of attributes. The lower bound was set up as at least 3 or more and the upper bound as unlimited, because more attributes yield better student measurements, according to many researchers. Equation (36) represents a constraint of discrimination values. The parameter r_{ik}^{*} is considered a discrimination value of item i for attribute *k*. It is the penalty for lacking attribute *k*. Therefore, it is better to have a small value when adding it across the items given each attribute. For this study r_{ik}^{*} value of 11 was used.

3.2.3 Maximin Method

The second method is called the Maximin Method (van der Linden & Boekkooi-Timminga, 1989):

Maximize y (39)

Subject to

$$\sum_{i=1}^{I} I_i(\theta_l)x_i - r_l y \geq 0 \qquad l = 1,\ldots, L \qquad (40)$$

$$\sum_{i=1}^{I} x_i = n \qquad i = 1, \ldots, I \qquad (41)$$

$$LB \leq \sum_{i=1}^{I} q_a x_i \leq UB \qquad a = 1, \ldots, A \qquad (42)$$

$$LB \leq \sum_{i=1}^{N} r_{ik}^{*} x_i \leq UB \qquad k = 1,\ldots, K \qquad (43)$$

and

$$x_i \in \{0,1\}, \qquad i = 1, \ldots N \qquad (44)$$

$$y \geq 0\,. \qquad (45)$$

All the constraints for the Maximin Method were the same as in the Minimax Method except Constraint (40), which shows that the relative target information function is characterized by a series of lower bounds (R_1y, …, R_ly). Therefore, this method, in a sense, maximizes the test information. The value of relative target information in this dissertation is 1, same as in the Minimax Method. Constraint (42) shows how to set lower and upper bounds for the number of attributes for each item. This constraint is similar to the content area balancing. Here, *A* stands for the number of attributes. Constraint (43) is related to the discrimination values. The parameter r_{ik}^{*} is considered a discrimination value of item i for attribute *k*. For this study, r_{ik}^{*} of 11 was used.

.

3.2.4 Maximum Information Method

The third method used, the Maximum Information Method, simply maximizes the test information function at θ_c:

Maximize

$$\sum_{i=1}^{I} I_i(\theta_c)x_i \qquad (46)$$

Subject to

$$\sum_{i=1}^{I} x_i = n \qquad i = 1, \ldots, I \qquad (47)$$

$$LB \le \sum_{i=1}^{I} q_a x_i \le UB \qquad a = 1, \ldots, A \qquad (48)$$

$$LB \le \sum_{i=1}^{N} r_{ik}^{*} x_i \le UB \qquad k = 1, \ldots, K \qquad (49)$$

and

$$x_i \in \{0,1\}. \qquad i = 1, \ldots, N \qquad (50)$$

The Maximum Information Method is similar to the previous two methods. One difference is that the objective function (46) has its maximum at the θ_c point, which was equal to -0.3 in this study.

To solve for the value of x_i (i = 1, …, I) and y, a branch-and-bound algorithm of 0/1 Linear Programming can be used. Currently, such an algorithm is available in computer code or in commercial linear programming packages such as CPLEX. In this dissertation GAMS software (Boisvert et al., 1985) was used to automatically select the best possible items. This software program can be used to find the best solution to the optimization function under the given constraints; however, the source code is not available to the public. The procedures of selecting items and analyzing the results are explained thoroughly below.

3.3 DATA GENERATION

Test Construction

Mimicking van der Linden (in press) and van der Linden and Boekkooi-Timminga (1989), this study considered three methods in the automated test assembly; Minimax, Maximin, and Maximum Information (van der Linden, in press; van der Linden & Boekkooi-Timminga, 1989). As explained above, each method provides specific objective functions with IRT-based constraints. Each of the parallel tests are constructed from three objective functions and a list of constraints using the commercial program GAMS (Boisvert et al., 1985) with CPLEX solver (ILOG, 2003). Then, the new constraints are added to the conventional IRT constraints in order to express the Cognitive Diagnostic aspect of the procedure (such as an assembled test Q-matrix, and information related to the discriminant).

After a certain number of parallel tests is assembled using 0/1 Linear Programming, each of the parallel tests is "administered" to the generated examinees. Then, these sets of items are used to generate hypothetical response sets from different collections of examinees.

After both of the items and persons parameters are calibrated, further analysis are performed to obtain a set of values describing the mastery level for each attribute as well as a traditional IRT ability estimate for each examinee.

IRT-Based Analysis

For a further comparison of parameters after the test assembly, ability parameters obtained using a Fortran program are treated as the true or known values, denoted as θ_0. The examinees' true θ levels are generated such that they later correlate with the estimated θ. The estimated θ are calculated once the test selected using GAMS software (Boisvert et al., 1985) with CPLEX solver is given to the true simulees.

The item response patterns are simulated by comparing the probability of acquiring a correct response. This probability can be compared with a number that is randomly drawn from a uniform distribution between 0 and unity. The probability of getting an item right given the examinee's true ability level (θ_0) is calculated from the 3PL model. If the probability of obtaining a correct response is greater than the random number, the item is said to be correct (that is, coded as 1). Otherwise, the item is indicated as incorrect (that is, coded as 0).

Then, given these responses, new estimates of $\hat{\theta}_j$ for each examinee j are calculated using the Maximum Likelihood Estimation Procedure from the Fortran program.

Cognitive Diagnostic-Based Analysis

From the above analysis, the automated test assembly methods based on IRT models can be examined. In this section, the way of analyzing the Cognitive Diagnosis models are explained. The attribute mastery parameters of each examinee are estimated

as continuous, and then they are dichotomized by evaluating their values at the cut-off value of 0.5. A value greater than or equal to 0.5 is assigned mastery status of the given attribute, and a value less than 0.5 is assigned non-mastery status for the given attribute. This mastery/non-mastery status is what is reported in the diagnostic score reports.

Similar to the single-ability estimate example, the vector of ability parameters obtained using the Arpeggio program are treated as the true attribute mastery patterns, denoted as $\underline{\alpha}_0$. The true α vectors are generated to compare with the estimated α vectors. These estimated α vectors are calculated after the test selected using GAMS (Boisvert et al., 1985) with CPLEX solver is given to the true simulees. New estimates of the Fusion Model parameter $\hat{\underline{\alpha}}_j$ for each examinee *j* are estimated by MCMC using Arpeggio programs. These new estimates of $\hat{\underline{\alpha}}_j$ are compared with known values, such as the true abilities parameters $\underline{\alpha}_0$.

To summarize how the item and person parameters are obtained for both the IRT-based analysis and the Cognitive Diagnostic-based analysis, refer to Figure 4.

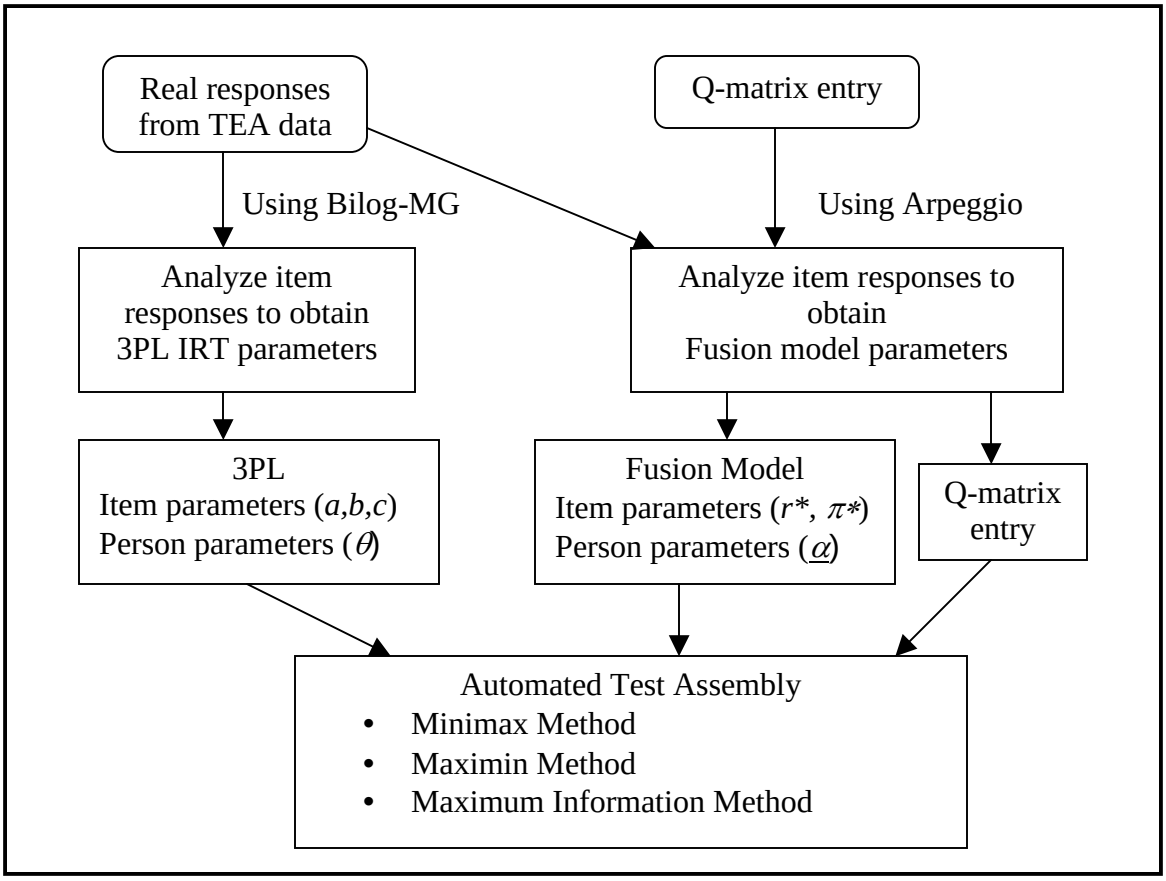


Figure 4: Obtaining IRT and Cognitive Diagnostic Theory parameters

While this study focuses on a particular Cognitive Diagnosis model, which is a Fusion Model, the approach can be generalized to any diagnostic model that involves the estimation of the examinees' knowledge states.

3.4 EVALUATION CRITERIA

Before evaluating the accuracy of tests that are constructed, first, it is needed to examine whether or not the parallel tests are constructed. To do that, the correlation between test information of first and second tests were calculated. Because the objective functions for all three methods are related to the information, test informations were examined to see whether the items are selected properly. This is essential to inspect whether two parallel tests were generated

It is important to evaluate whether the attribute mastery level estimates and the traditional IRT ability estimates are both accurate. It is also necessary to assess how well the automated test assembly methods select the appropriate items. To do this, three criteria are used in subsequent sections to compare the results from the simulation study.

To evaluate whether the given condition worked well, IRT ability estimates $\hat{\theta}_j$ (from the automatically generated tests) are compared with the corresponding values of the ability parameter obtained from the Fortran Program of the original dataset θ_{j0} for each examinee *j*. Likewise, evaluation of the attribute mastery level is carried out by comparing the estimates of $\hat{\underline{\alpha}}_j$ (from the automatically generated tests) with the attribute vectors from the Arpeggio analysis of the real data $\underline{\alpha}_0$ for each examinee *j*. If a given method works well, then the new estimates of both parameters θ_0 and $\underline{\alpha}_0$ should be equivalent to (or similar to) the corresponding true parameters θ and $\underline{\alpha}$ (McGlohen, 2002).

First, the correlations are calculated between the estimated values $\hat{\theta}_j$ and the true values θ_0 from the real data. If these correlations are high, then the methods are working well. In addition, the comparison between the true theta θ_0 and its corresponding estimate $\hat{\theta}_j$ are determined by examining the root-mean-square error (RMSE), the mean-square error (MSE) and the bias statistics.

Second, the information function plots for all the parallel tests are compared by means of visual inspections. After all the parallel tests have been constructed, their information functions will be calculated and then plotted on the same scale. Then, the differences between the information functions are compared.

Third, the hit rates for attribute mastery are calculated between the estimated values and the true values from the real data. The estimates of $\hat{\underline{\alpha}}_j$ are compared to the true attribute vectors from the original real data provided by the Arpeggio analysis. This procedure allows examination of the hit rate for each attribute as well as the hit rate for the entire attribute pattern for each examinee (McGlohen, 2004).

Also, the proportion of flagged examinees is calculated using the additional software in Arpeggio package. The flagged examinee can be described as two different kinds, one with a low probability of achieving mastery status while indeed obtaining mastery status and other with a high probability of achieving mastery status while who did not obtain mastery status. The automated assembly method with the highest correlations and hit rates is considered to be assembled best compared to the remaining methods.

Using these criteria, results can be evaluated with respect to the accuracy of both the attribute classification rate and the ability levels by comparing the estimated values with the simulated true values.

CHAPTER 4: RESULTS

Three criteria were used to evaluate the results of this dissertation. For all three criteria, two sets of results are presented for two parallel tests.

First, the IRT-based analysis was done to evaluate the already existing IRT-based automated test assembly methods. It was essential that the various methods accurately estimate the values of the single score, $\hat{\theta}$. Therefore, the true values of θ_0 were compared with the values of $\hat{\theta}_j$. The correlations between these two values were then calculated. The comparison between the true theta θ_0 and its corresponding estimate $\hat{\theta}_j$ were determined by examining the root-mean-square error (RMSE), the mean-square error (MSE) and the bias statistics.

Second, the items had to be selected based on the shape of the target information functions. The visual inspections of test information were used to evaluate the selection of the items. In addition, correlations of test information between first and second tests were calculated to check the parallel tests.

Third, the cognitive diagnostic-based analysis was conducted to evaluate how well the newly developed constraints are working. The means and standard deviation of π_i* estimates were presented for the Minimax, Maximin, and Maximum Information Methods. Also, the means and standard deviations of *r** parameters were described for the three automated test assembly methods.

In addition, each method had to accurately estimate the attribute mastery patterns. Consequently, the correct classification rates were calculated for each measured attribute

as well as for the entire attribute pattern. The evaluation of the attribute mastery level was carried out by comparing the estimates of $\hat{\underline{\alpha}}_j$ (from the automatically generated tests) with the attribute vectors from the Arpeggio analysis of the real data $\underline{\alpha}_0$ for each examinee *j*.

4.1 IRT-BASED ANALYSIS

Nonconvergent cases for each condition were removed from the analyses. Table 1 illustrates the number of nonconvergent cases for each condition. Non-convergence was decided by the estimate being assigned out of ranges for -4 or 4. The smallest of 23 to the largest of 129 non-convergent cases were found. Among the three conditions, the Discrimination-only Condition showed the largest values of nonconvergent cases.

Table 1: Nonconvergent cases for three different test construction methods

	Minimax		Maximin		Max Info	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Attribute Only	41	23	24	60	104	26
Discrimination Only	129	73	39	73	120	76
Both Attribute & Discrimination	39	44	57	39	59	34

As described above, the theta estimates of the different methods were of particular interest. The true thetas were generated from normal distribution. These true theta values were compared with the theta estimates to verify whether each of the methods was successful in accurately estimating the single score theta. The comparison between the true theta and its corresponding estimate was done by examining the values of the correlation coefficient between two values, the root-mean-square error (RMSE), the mean-square error (MSE), and the bias statistics.

Correlation coefficients calculated between the true θ_0 and its corresponding estimate $\hat{\theta}_j$ are shown in Table 2 for the response probabilities based on three different test construction methods. The RMSE values are presented in Table 3 for probabilities based on the three different test construction methods. The MSE and bias statistics are shown in Tables 4 and 5, respectively.

Table 2: Correlations of the true theta values and the estimated theta values for three different test construction methods

	Minimax		Maximin		Max Info	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Attribute Only	0.914	0.858	0.958	0.905	0.949	0.923
Discrimination Only	0.901	0.843	0.954	0.888	0.938	0.927
Both Attribute & Discrimination	0.941	0.847	0.940	0.924	0.944	0.897

Table 3: Root Mean Square Error of the estimated theta values for three different test construction methods

	Minimax		Maximin		Max Info	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Attribute Only	0.422	1.091	0.298	0.604	0.273	0.678
Discrimination Only	0.422	1.253	0.296	0.602	0.291	0.513
Both Attribute & Discrimination	0.346	1.157	0.338	0.634	0.321	0.589

Table 4: Mean Square Error of the estimated theta values for three different test construction methods

	Minimax		Maximin		Max Info	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Attribute Only	0.178	1.189	0.088	0.365	0.074	0.460
Discrimination Only	0.143	1.253	0.087	0.363	0.085	0.263
Both Attribute & Discrimination	0.120	1.157	0.114	0.402	0.103	0.347

Table 5: Bias statistics of the estimated theta values for three different test construction methods

	Minimax		Maximin		Max Info	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Attribute Only	0.042	-0.959	-0.009	-0.443	-0.001	-0.559
Discrimination Only	-0.008	-1.106	0.002	-0.409	-0.011	-0.368
Both Attribute & Discrimination	-0.0004	-1.022	-0.01	-0.509	0.017	-0.411

In Table 3, the correlations between the true thetas and estimated thetas are moderate. Under each condition, different results were yielded. All the correlation values are above 0.84. Moreover, the Maximin method and Maximum Information method of the first test show fairly high correlations, from 0.94 to 0.96. Correlations are typically lower for the second generated test than for the first test. The range of correlations for the second test was from 0.84 to 0.93. In general, the first test tends to have more accurate estimates than the second test.

In Table 3, the root-mean-square error is lower for the first exam where the item selections are based on the full item pool. Also, the values of MSE and bias statistics are pretty small for the first test of all three methods. The values increased slightly more for the second test. Overall, the conditions of the Maximum Information Method seem to perform comparably at accurately measuring the single score $\hat{\theta}_j$ for the examinees. The Maximin Method perform comparably well at estimating $\hat{\theta}_j$ for the examinees as well. By examining the values of RMSE, MSE, and bias statistics, it can be seen that the Minimax Method is not performing accurately in the case of the second test.

The first test shows higher correlation, lower RMSE values and MSE values, and the lower bias statistics demonstrate that the test generated first based on item selection of the most optimal condition performs better than when the items were selected after optimal items were excluded.

4.2 ITEM INFORMATION ANALYSIS

For the second criterion, the test information functions from the two tests were compared to target test information. All of the test information curves show appropriate shapes to demonstrate that items were properly selected. Test information for each of the methods is plotted in Appendix C.

For the Minimax Method, absolute targets were used to select the items (Figures 5 though 7). When absolute targets are used to assemble tests, a fixed number of information units are needed at the θ_l points. The absolute target values for the TIF at θ_l, l = -1.0, 0, +1.0 are 8.25, 8.33, 3.32, respectively. These values are based on the actual test from TEA. The means of test information were calculated for the administration of three math tests. From Figures 5 through 7, it can be verified that no test information exceeded the value of 10. For all three conditions, the first test fits the absolute target information values better than the second test.

For the Maximin Method, the relative target values for the TIF at θ_l, l = -1, 0; 0,; and +1.0 are 1, 1, and 1, respectively. These relative target values were used for the simulated test construction in van der Linden (1998)'s article. The shapes of all three test information are symmetric in Figures 8 through 10. The first test seems to be more symmetrical than the second test. .

For the Maximum Information Method, the items were selected to maximize the information at θ_l = -0.3. As can be seen in Figures 11 and 13, the information peaked

when θ = -0.3. Except for Both Attribute and Discrimination Constraint, the information of the first test peaked higher than the second test.

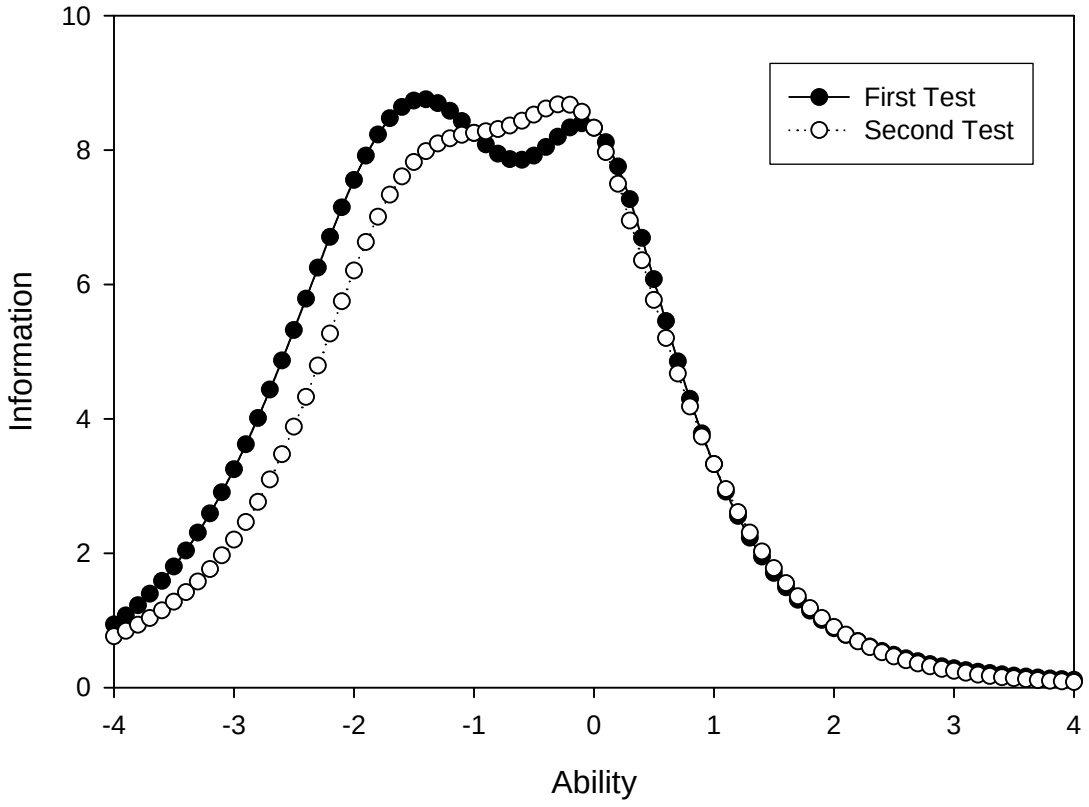


Figure 5: Test Information of First and Second Tests for the Minimax Method: Attribute-only Constraint

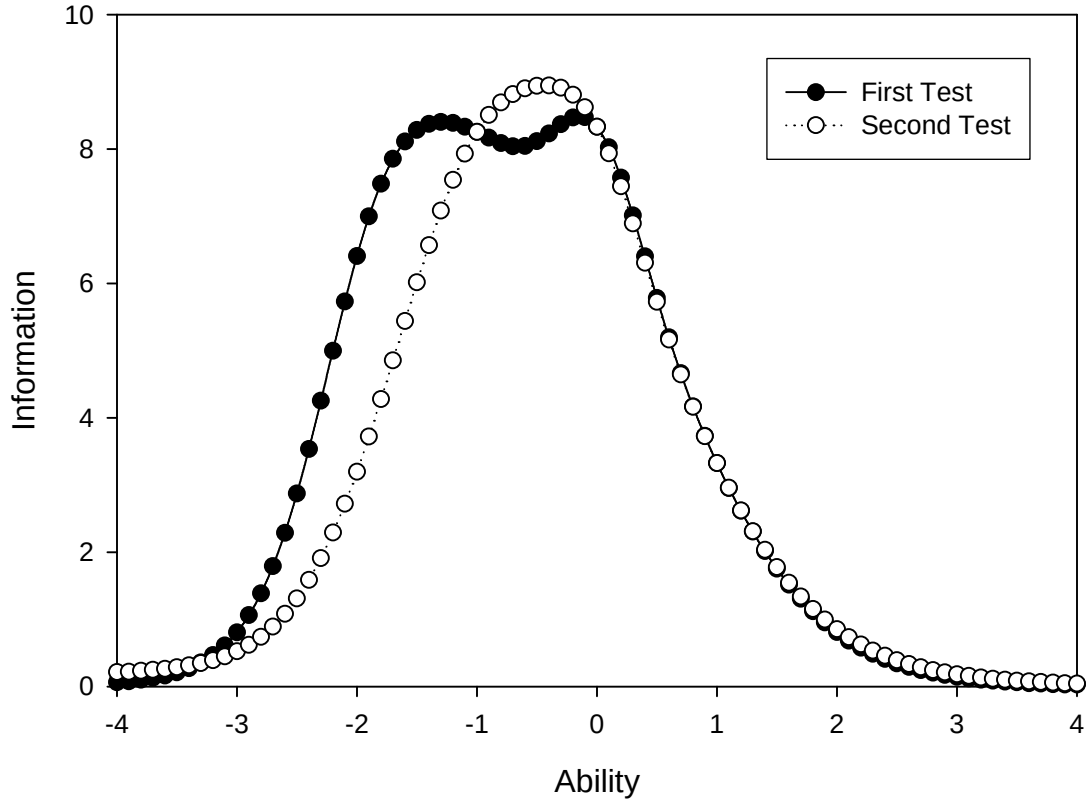


Figure 6: Test Information of First and Second Tests for the Minimax Method: Discrimination-only Constraint

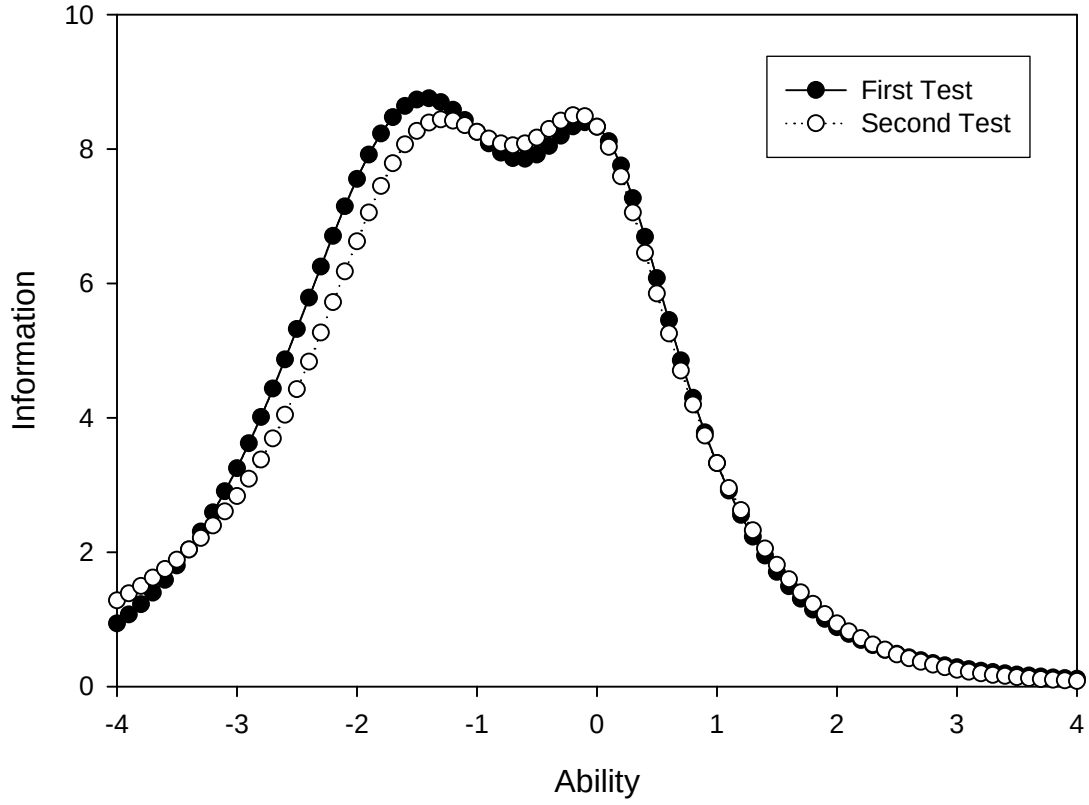


Figure 7: Test Information of First and Second Tests for the Minimax Method: Both Attribute and Discrimination Constraint

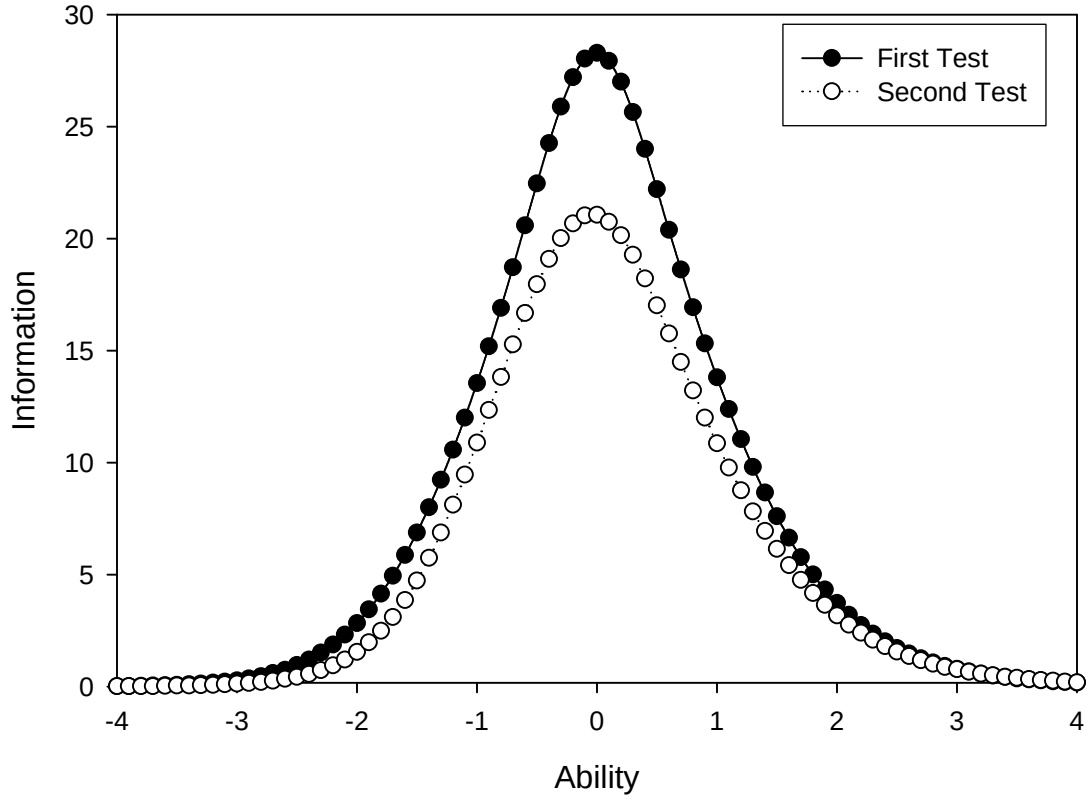


Figure 8: Test Information of First and Second Tests for the Maximin Method: Attribute-only Constraint

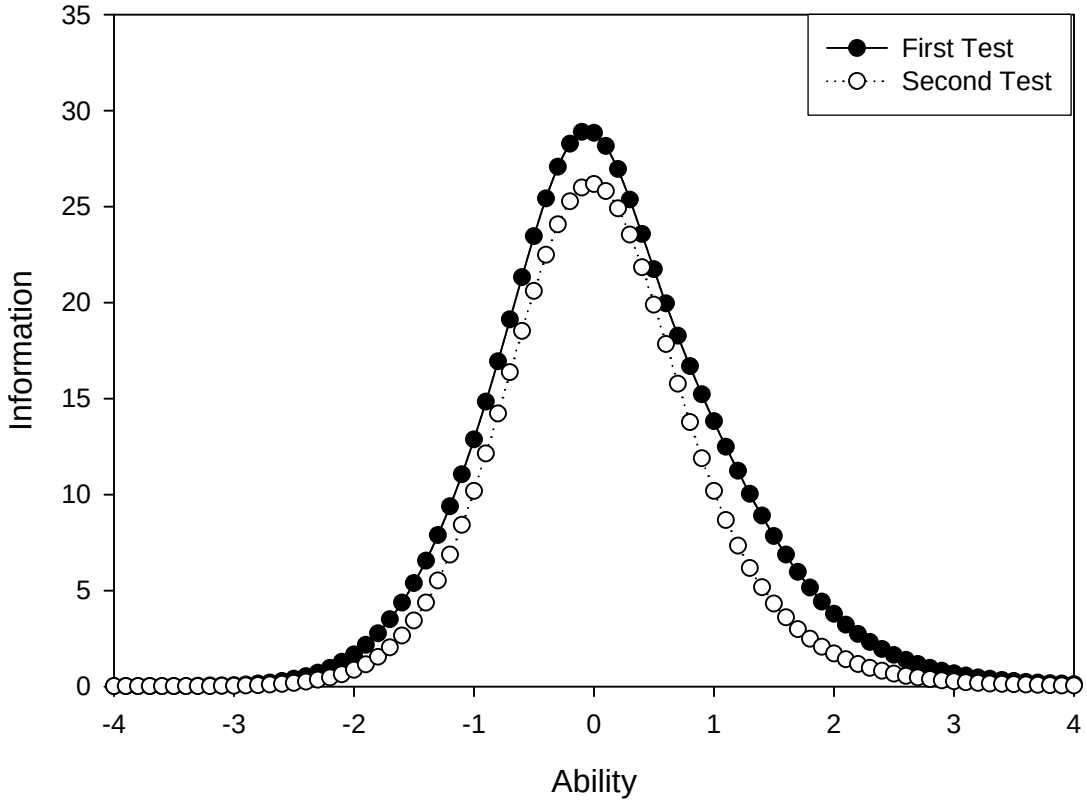


Figure 9: Test Information of First and Second Tests for the Maximin Method: Discrimination-only Constraint

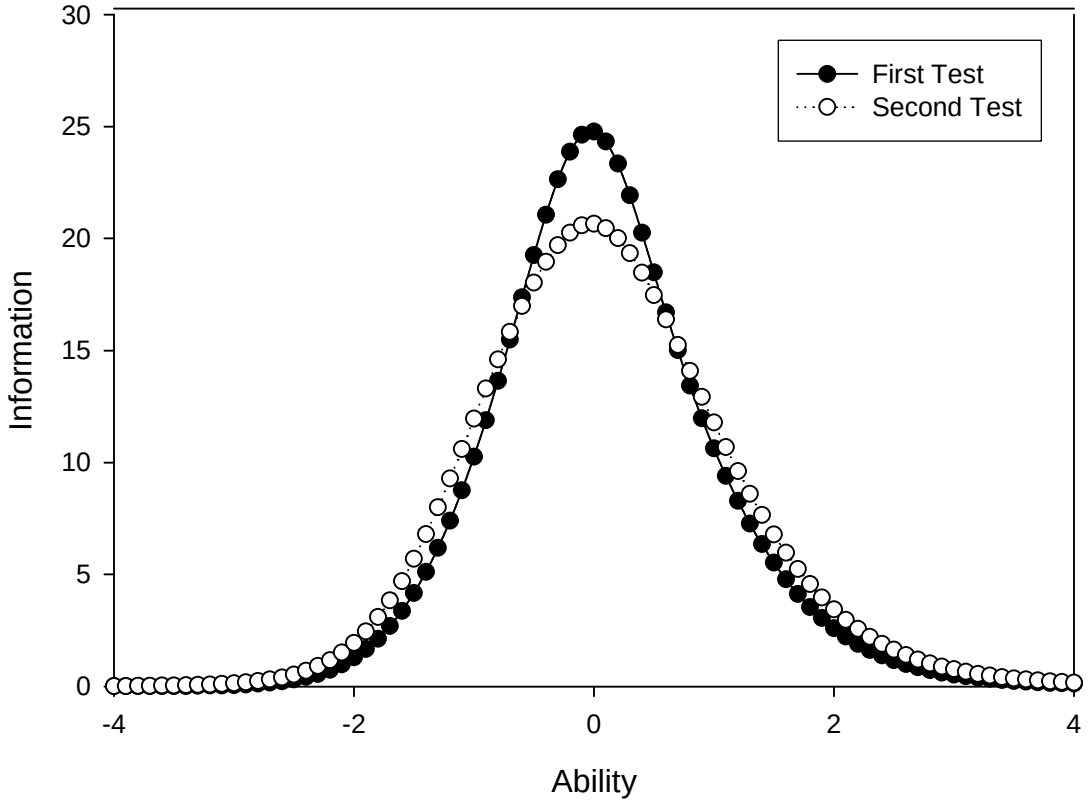


Figure 10: Test Information of First and Second Tests for the Maximin Method: Both Attribute and Discrimination Constraint

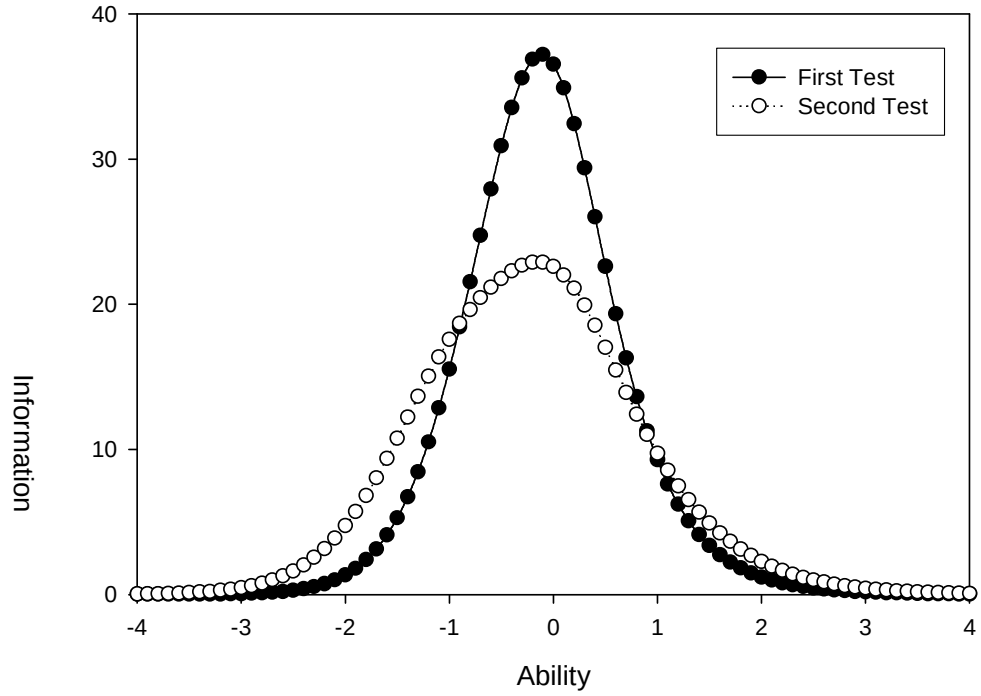


Figure 11: Test Information of First and Second Tests for the Maximum Information Method: Attribute only Constraint

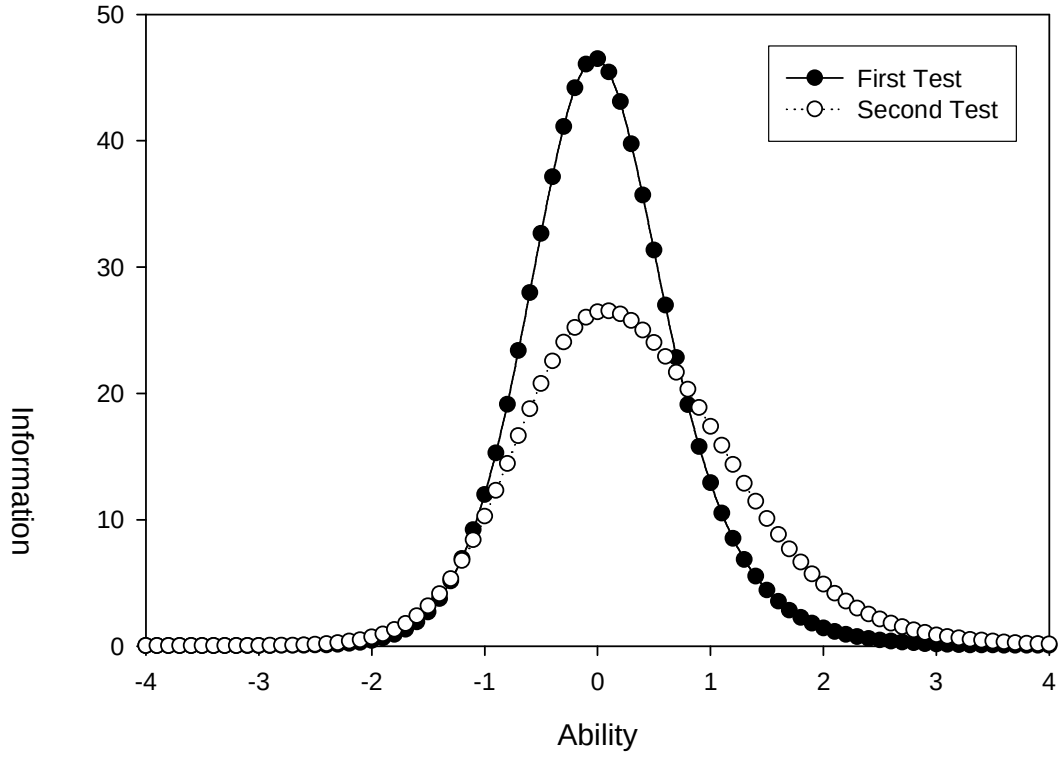


Figure 12: Test Information of First and Second Tests for the Maximum Information Method: Discrimination-only Constraint

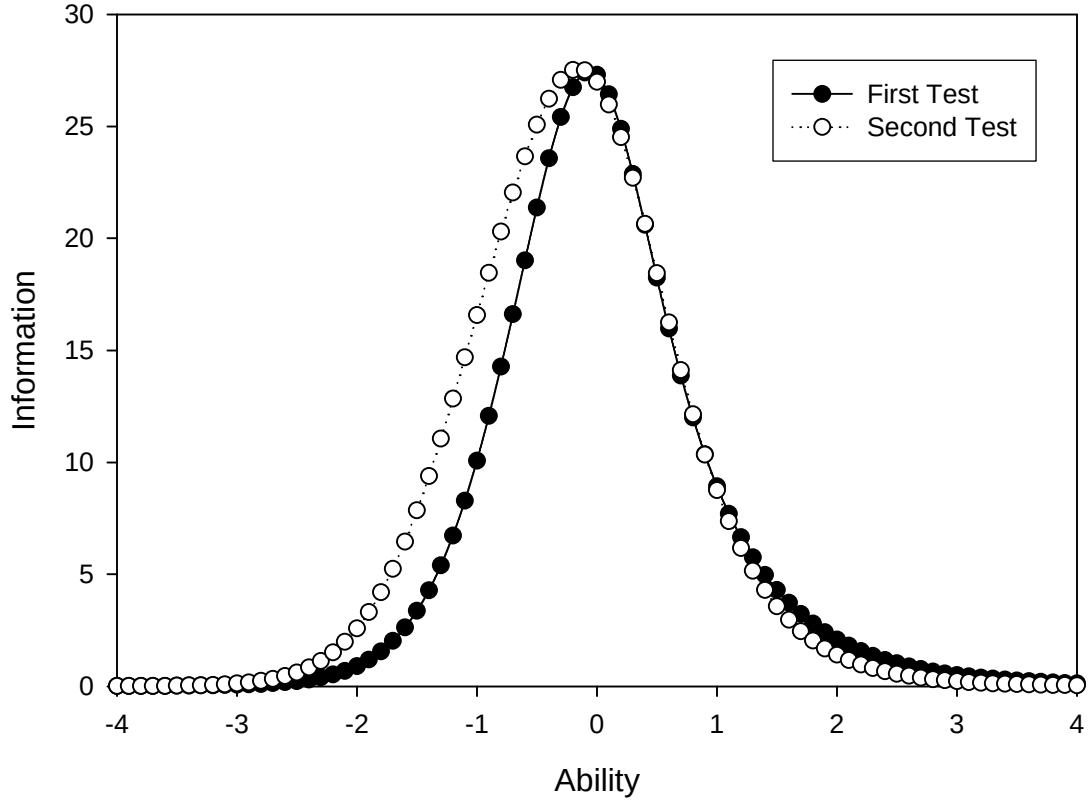


Figure 13: Test Information of First and Second Tests for the Maximum Information Method: Both Attribute and Discrimination Constraint

Because the objective functions for all three methods are related to the information, test information was examined to see whether the items were selected properly. The correlation between information of the first and second tests was calculated. This value was essential to inspecting whether the two parallel tests were generated. Table 6 shows a very high correlation between the two tests. Moreover, all the correlations are higher than 0.94, with the Maximin Method showing especially high correlation. For the Minimax and Maximin Methods, the correlation of test information was the highest when only the Attribute-only Constraint was added, as opposed to the other condition. However, for the Maximum Information Method, the correlation was the highest when Both Attribute and Discrimination Constraints were added.

Table 6: Correlation of test information between two parallel tests: first test and second test

	Minimax	Maximin	Max Info
Attribute Only	0.985	0.999	0.953
Discrimination Only	0.954	0.994	0.947
Both Attribute & Discrimination	0.965	0.989	0.975

4.3 COGNITIVE DIAGNOSTIC-BASED ANALYSIS

First, the numbers for each attribute are compared for each of the three conditions in all three methods. Appendix D shows how many attributes are included in the test. It also shows which attributes are not selected during item selection, as well as the conditions for that nonselection. The second constraint, which is the discrimination-only constraint, cannot select all the attributes in the Q-matrix. In Appendix D, the Maximin Method with discrimination-only constraint did not select items that contain the attributes 2, 3, and 5. Instead, this method selected items with a large amount of attribute 10.

To evaluate whether the test was assembled appropriately according to Cognitive Diagnosis theory, the Q-matrix of each automated test assembly methods was examined. The analysis of the Q-matrix is shown below. Tables 7 through 9 represent the means and standard deviation of π_i^* estimates for the Minimax, Maximin, and Maximum Information Methods, respectively. The parameters of π_i^* are known as the probability of correctly applying all item *i* required attributes when the examinee has mastered all attributes required by item *i* (Hartz et. al, 2002).

The high values of π_i^* indicate that the item is a good item. As shown below, most of the means of π_i^* show the high values ranges from 0.79 to 0.91.

Table 7: Means and standard deviations of π^* estimates for the Minimax Method.

	Minimax		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.868	0.858	0.786
	(0.040)	(0.043)	(0.059)
Test 2	0.885	0.861	0.891
	(0.034)	(0.035)	(0.033)

Table 8: Means and standard deviations of π^* estimates for the Maximin Method.

	Maximin		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.831	0.817	0.839
	(0.036)	(0.037)	(0.047)
Test 2	0.846	0.863	0.839
	(0.039)	(0.050)	(0.037)

Table 9: Means and standard deviations of π^* estimates for the Maximum Information Method.

	Max Info		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.907	0.894	0.853
	(0.041)	(0.041)	(0.055)
Test 2	0.878	0.805	0.902
	(0.038)	(0.037)	.(0.038)

Table 10 through 12 also present the means and standard deviations of r^* parameters for the three automated test assembly methods, respectively. To be a good item, the values of r^* need to be low. Items containing high r^* values for a certain attribute might not measure that attribute well. The values of r^* should remain less than 0.9. According to the Arpeggio manual (Hartz et al, 2002), the estimated r^* values of this study were in the acceptable range. The mean r^* values are generally moderate values (between 0.41 to 0.59) except for the Minimax Method. For the Minimax Method, the mean r^* values for Attribute-only constraint were 0.70 and 0.81 for test 1 and test 2, respectively. Also, the mean r^* value was 0.71 for the Attribute & Discrimination Constraint in test 2; however, the value was very small, 0.41 for test 1.

When the items are considered to be included in the test, low values of r^*'s and high values of π^*'s are preferred. These parameters of r^* and π^* can be employed to evaluate the performance of each items in the test. Then, poorly performing items can be examined. This procedure can be a good way of evaluating how well the tests are constructed based on Cognitive Diagnosis models.

Table 10: Means and standard deviations of *r** estimates for the Minimax Method.

	Minimax		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.698	0.580	0.410
	(0.079)	(0.089)	(0.069)
Test 2	0.656	0.525	0.709
	(0.085)	(0.067)	(0.085)

Table 11: Means and standard deviations of *r** estimates for the Maximin Method.

	Maximin		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.513	0.474	0.479
	(0.064)	(0.049)	(0.065)
Test 2	0.580	0.502	0.575
	(0.072)	(0.070)	(0.067)

Table 12: Means and standard deviations of r^* estimates for the Maximum Information Method.

	Max Info		
	Attribute	Discrimination	Attribute & Discrimination
Test 1	0.548	0.482	0.485
	(0.072)	(0.062)	(0.067)
Test 2	0.590	0.505	0.582
	(0.068)	(0.060)	(0.059)

To evaluate the attribute-mastery estimation, the correct classification rates of each measured attribute and the entire attribute pattern are presented in Tables 13 through 15. The correct classification rates are also called "hit rates." Table 13 presents the correct classification rates of each test using the Minimax Method to determine the response pattern probabilities for the math test. Tables 14 and 15 present the correct classification rates of each test using the Maximin and Maximum Information Methods.

For all three methods, the second test classifies the examinees more correctly than the first test for the Attribute-only Constraint and for Both Attribute and Discrimination Constraints. The Discrimination-only constraint yields slightly lower correct classifications for the second test than the first test.

For the Maximin Method and the Maximum Information Method, the Attribute-only Constraint correctly classifies the examinees more consistently as masters or non-masters of the measured attributes, while Discrimination-only constraint shows more fluctuation. The Minimax Method, however, shows a different pattern. In this case, discrimination-only constraint seems to generate more accurate attribute mastery classifications for the first test. For the second test, all three conditions seem to yield similar results, but the Attribute-only Constraint classifies the examinee more accurately. Comparisons of each attribute are irregular across the three methods.

As can be seen from the standard deviation, it would be preferable to use Both Attribute and Discrimination constraints. The standard deviations of Both Attribute and Discrimination constraints are ranged from 0.053 to 0.072. However, the standard deviations of first test for the Minimax Method are 0.294 and 0.139 (Attribute-only and Discrimination-only constraints), respectively. Also, the standard deviation of the

Discrimination-only constraint on the second test for the Maximin Method is 0.106. These indicate that using either one separately is not stable. Both of the diagnostic constraints together make the result more consistent.

Table 13: The math test's attribute mastery hit rates for the Minimax Method for both tests

	First Test			Second Test		
Attribute	Attr only	Discrim.only	Both	Attr only	Discrim.only	Both
1	0.134	0.583	0.423	0.818	0.826	0.714
2	0.085	0.475	0.533	0.773	0.664	0.797
3	0.182	0.918	0.549	0.880	0.642	0.892
4	0.297	0.519	0.548	0.740	0.733	0.766
5	0.291	0.844	0.488	0.784	0.803	0.822
6	0.395	0.702	0.654	0.722	0.714	0.796
7	0.594	0.763	0.607	0.763	0.688	0.760
8	0.866	0.725	0.572	0.747	0.685	0.742
9	0.791	0.530	0.510	0.707	0.577	0.736
10	0.772	0.630	0.415	0.653	0.492	0.680
11	0.766	0.677	0.498	0.685	0.643	0.666
Mean 1-11	0.470	0.670	0.527	0.752	0.679	0.761
Std Dev	0.294	0.139	0.072	0.063	0.095	0.065

Table 14: The math test's attribute mastery hit rates for the Maximin Method for both tests

	First Test			Second Test		
Attribute	Attr only	Discrim.only	Both	Attr only	Discrim.only	Both
1	0.451	0.485	0.452	0.551	0.554	0.528
2	0.575	0.611	0.546	0.620	0.416	0.623
3	0.743	0.546	0.648	0.623	0.475	0.621
4	0.551	0.567	0.576	0.557	0.464	0.540
5	0.500	0.613	0.498	0.612	0.297	0.589
6	0.668	0.715	0.700	0.681	0.664	0.650
7	0.600	0.618	0.580	0.649	0.645	0.618
8	0.616	0.650	0.597	0.585	0.531	0.572
9	0.516	0.546	0.528	0.519	0.542	0.514
10	0.464	0.485	0.462	0.401	0.506	0.458
11	0.544	0.548	0.541	0.513	0.602	0.520
Mean 1-11	0.566	0.580	0.557	0.574	0.518	0.567
Std. Dev.	0.088	0.069	0.075	0.078	0.106	0.059

Table 15: The math test's attribute mastery hit rates for the Maximum Information Method for both tests

	First Test			Second Test		
Attribute	Attr only	Discrim.only	Both	Attr only	Discrim.only	Both
1	0.573	0.563	0.464	0.495	0.474	0.554
2	0.578	0.512	0.576	0.709	0.497	0.625
3	0.665	0.536	0.601	0.754	0.513	0.678
4	0.584	0.577	0.576	0.695	0.601	0.608
5	0.636	0.517	0.481	0.695	0.478	0.647
6	0.676	0.515	0.665	0.699	0.485	0.698
7	0.622	0.597	0.599	0.667	0.633	0.627
8	0.623	0.508	0.619	0.649	0.582	0.637
9	0.533	0.517	0.512	0.525	0.520	0.540
10	0.567	0.556	0.558	0.554	0.460	0.556
11	0.537	0.524	0.537	0.555	0.544	0.558
Mean 1-11	0.599	0.538	0.563	0.636	0.526	0.612
Std. Dev.	0.048	0.030	0.060	0.088	0.057	0.053

Another way of evaluating how well the automated test assembly methods are working is to examine the flagged examinees. The proportion of flagged examinees should be small to ensure that the test is soundly constructed. Table 16 represents the proportion of flagged examinees for each method. There should be some people with attribute-mastery estimate values between 0.4 to 0.6. Even though these values are very close to the cutoff value of 0.5, these people can be misclassified. Therefore, the smallest possible number of these people is preferred. All three methods show less than 10 % or slightly more than 10% of flagged examinees except the Minimax 1 Attribute-only condition.

Table 15: Proportion of flagged examinees

	Attribute	Discrimination	Attribute & Discrimination
Minimax			
Test 1	0.583	0.150	0.121
Test 2	0.177	0.093	0.193
Maximin			
Test 1	0.058	0.070	0.048
Test 2	0.091	0.162	0.116
Max Info			
Test 1	0.124	0.182	0.061
Test 2	0.055	0.196	0.141

4.4 OVERALL PERFORMANCE

So far, the automated test assembly methods of incorporating both IRT and Cognitive Diagnostic Model were evaluated using three criteria.

Overall, two parallel tests were constructed automatically, and the correlation of information between the two tests was moderately high, indicating that the tests were parallel. This result can be validated also from Figures 5 to 13. The IRT-based analysis shows that all three automated test assembly methods result in better θ estimation. The first test generated, however, seemed to have more accurate θ estimation than the second test generated. This is due to the item selection. The first test selects the best items from the item pool creating the better test information function, then the second test was generated.

For the Cognitive-Diagnostic-based analysis, the test was created using three methods that fit the Cognitive Diagnostics models well. The Q-matrix, which comprises selected items, showed that appropriate items were picked; however, the Discrimination-only constraint condition did not pick certain attributes. It can be concluded, therefore, that both of the diagnostic constraints needed to be added to generate good parallel tests.

The results of the Cognitive-Diagnostic-based analysis were not as clear and consistent as the result of the IRT-based analysis. The reasons for this difference are that the Q-matrix is overly simple and that the upper and lower limits of the constraints are unclear. These reasons are explained in the Limitations and Future Research chapter.

CHAPTER 5: DISCUSSIONS

This chapter summarizes the study and its overall results and then considers the significance of the study in the field of measurement. Finally, the chapter describes the limitations of the study as well as its implications for future research.

5.1 SUMMARY AND COMMENTS

The commonly used automated test assembly methods involve a variety of test specifications, including content balancing, item format, section length, test length, reliabilities, count of words, and many more (van der Linden, 1998). Even though combinations of these constraints have been successful, the problems of constructing tests with Cognitive Diagnostic constraints on item selection have only recently been addressed. This dissertation has discussed how automated test assembly can be used in conjunction with the Cognitive Diagnosis framework to generate effective and beneficial test development methods. In particular, the study incorporated Cognitive Diagnosis elements into the IRT-based automated test assembly methods that satisfy requirements in both fields.

Three IRT-based 0/1 Linear Programming methods were used to address the different objective functions: the Minimax Method, Maximin Method, and the Maximum Information Method. Based on these three methods, aspects of the Cognitive Diagnosis model were incorporated to assemble tests automatically. In addition, some of the constraints related to Cognitive Diagnosis Theory were added to the currently existing automated test assembly methods that are based on IRT.

To select items automatically, three steps had to be identified. First was identification of the objective function to be optimized (for example, by maximizing test information or minimizing the sum of the positive deviations from the target test information). Second was the formulation of constraints for conventional test specifications (such as test length and contents). Third was the addition of the new, Cognitive Diagnostic constraints (such as an assembled test Q-matrix, and information related to the discriminant), which were combined with the conventional IRT constraints. All three steps were required to achieve optimization.

Two parallel tests were constructed automatically, and the correlation of information between the two tests was moderately high, indicating that the tests were parallel. The IRT-based analysis shows that three automated test assembly methods result in better θ estimation. The first test generated, however, seemed to have more accurate θ estimation than the second test generated.

For the Cognitive-Diagnostic-based analysis, the test was created using three methods that fit the Cognitive Diagnostics models well. However, three methods show the below average mastery classification rates just looking at the mean values. Mastery classification rates of second parallel tests were higher than the ones of first tests. This result provided the evidence that more diagnostic constraints need to be added to generate good parallel tests. Both of the diagnostic constraints were developed to incorporate Cognitive Diagnosis into IRT. Attribute constraint of constraining the Q-matrix, which comprises selected items, showed that appropriate items were picked; however, the discrimination-only constraint condition did not pick certain attributes. The analysis of the attribute mastery hit rates indicates that more specified diagnostic constraints are

needed to show consistent results. Thus, further studies are needed to develop better diagnostic constraints that can generate good cognitive diagnostic tests.

5.2 THE IMPORTANCE OF THIS STUDY

Conceivably, the automated test assembly methods described in this dissertation could gain wide use. First, the Cognitive Diagnostic approach, when used in conjunction with traditional standardized test assembly methods, gives testing specialists an algorithm that facilitates the development of tests. Second, the methods can easily be integrated into a real educational setting to provide cognitive diagnostic information. Third, the resulting diagnostic, information, which is closely associated with formative assessments, can prove much more helpful to learners and educators alike than any single score.

Finally, the approach yields diagnostic information from large-scale assessments and thus satisfies a requirement of the No Child Left Behind Act of 2001. Since the passage of this act, educational assessment has increased in importance in certain measured-content domains. Therefore, it is essential to acquire effective techniques for incorporating Cognitive Diagnosis in test development. The issue now is to identify more effective and beneficial test development methods within Cognitive Diagnosis frameworks that use automated test assembly. The solution offered by this dissertation is the incorporation of Cognitive Diagnosis elements into IRT-based automated test assembly methods.

5.3 LIMITATIONS AND FUTURE RESEARCH

There were some limitations on the scope and methods of this study that future studies may consider to carry the research further.

First, because the Q-matrix was taken from a real assessment of TEA, all items measured only a single attribute. For other tests, more complex Q-matrices that yield quite different results might be useful to capture the necessary attributes. In fact, it would be interesting to compare the results of a complex-structured Q-matrix to those of a single-structured one.

Second, the parallel tests represented two non-overlapping item structures. Because a certain number of items were excluded from the second test, the first test seems to have a better θ estimation than the second test. Therefore, a certain constraint might be needed to ensure that the second test is as good as the first test. For future study, over-lapping item structures might be examined. To obtain a diverse set of items from the item pool, an exposure control mechanism needs to be considered. For example, a certain stratification method (Chang & Ying, 1996; Chang & Ying, 1999) might be imposed; otherwise, if an exposure control method is not used for the test assembly, the popular items will always be exposed while the unpopular items will remain in the pool.

Third, for this research, two newly developed constraints were added to the already existing IRT-based automated test assembly methods. These additions were necessary to simplify matters. Because there is little research examining the advantages of combining IRT-based and Cognitive Diagnostic test assembly methods, the purpose of

this study was to determine whether their combination was practicable. Further research should now focus on developing more constraints for the Cognitive Diagnostic aspects.

This study lays the foundation for a rich field of research. Future topics include the following.

First, simulation studies would be effective in establishing more practical upper and lower limits of r^* parameters by. Also crucial is some examination of the effects of changing constraints such as the number of items measuring each attribute. These could be related directly to the attribute-only constraint that was used in this dissertation.

In addition, it would be interesting to examine more attribute-related constraints, such as the number of attributes measured as well as the number of attributes measured by a single item. Finally, in this dissertation, only the attribute mastery estimates of the cut-off value were examined. It is also important to deal with attribute mastery estimates close to the cut-off value.

5.4 CONCLUSION

As large-scale testing increases in national importance, educators must be able to transform scores from standardized testing into skill-level "formative assessments," tools that can aid the teaching and learning process, rather than into simplistic, single score based "summative assessments." Formative assessments help students, parents, and teachers understand the students' intellectual strengths and weaknesses, which is more useful information than any single score for what might really be a complex set of

abilities. Cognitive Diagnosis modeling, with IRT modeling as a foundation, is one way to achieve formative assessments.

Even though research attention on Cognitive Diagnosis is growing, only a few studies have addressed procedures for assembling tests according to given cognitive criteria. This dissertation, however, has generated such a procedure. It was the purpose of this dissertation, after all, to develop a test-construction method using an automated item selection method to develop a test, and this procedure was based on the item-related properties and statistical principles of cognitive-diagnosis and IRT models. The formative assessment tests were constructed using the automated item selection method. Even though this research was based on the Fusion Model, the application can be generalized to any diagnostic model that estimates the attribute states of the examinees.

By combining the strengths of the conventional testing framework and the new capabilities of the Cognitive Diagnostic framework, this new method will benefit many fields in educational and psychological testing. Indeed, Cognitive Diagnostic assessments give both learners and educators the means to diagnose correctly the learners' knowledge states. Granted, the process of assembling tests based on Cognitive Diagnosis can be complex. The research described in this dissertation, however, is able to reduce the complexity by applying and improving available technologies that automate the task.

Appendix A

List of Fourteen Cognitive Diagnosis Models

Models	Authors	Estimating Examinee Attributes	Relating Items to Attributes	Statistically Identifiable Parameters
LLTM	Fischer (1973)	No	Yes	Yes
MLTM	Whitley (1980)	Yes	No	Yes
Rule Space	Tatsuoka & Tatsuoka (1982)	Yes	No	Yes
GLTM	Embretson (1984)	Yes	Yes	No
Binary Skills Model	Haertel (1989)	Yes	No	Yes
HYBRID	Gitomer & Yamamoto (1991)	Creates own cognitive structure		Yes
Unified Mode	DiBello, Stout & Roussos (1993)	Yes	Yes	No
Bayesian Networks	Mislevy (1994)	Yes	No	Yes
Tree Based Approach	Sheeha (1997)	Yes	No	Yes
Discrete Mixture Rasch Model	Bolt (1999)	Creates own cognitive structure		Yes
Conjunctive, Disjunctive, & Compensatory MCLCM	Maris (1999)	Yes	Yes	No
Dichotomization of MLTM	Junker (2001)	Yes	No	Yes

Appendix B

Attributes Measured by Math Test

1. Demonstrate an understanding of number of concepts.

2. Demonstrate an understanding of mathematical relations.

3. Demonstrate an understanding of geometric properties and relationships.

4. Demonstrate an understanding of measurement concepts using metric and customary units.

5. Demonstrate an understanding of probability and statistics.

6. Use the operation of addition to solve problems.

7. Use the operation of subtraction to solve problems.

8. Use the operation of multiplication and/or division to solve problems.

9. Estimate solutions to a problem situation and/or evaluate the reasonableness of a solution to a problem situation.

10. Determine solution strategies and analyze or solve problems.

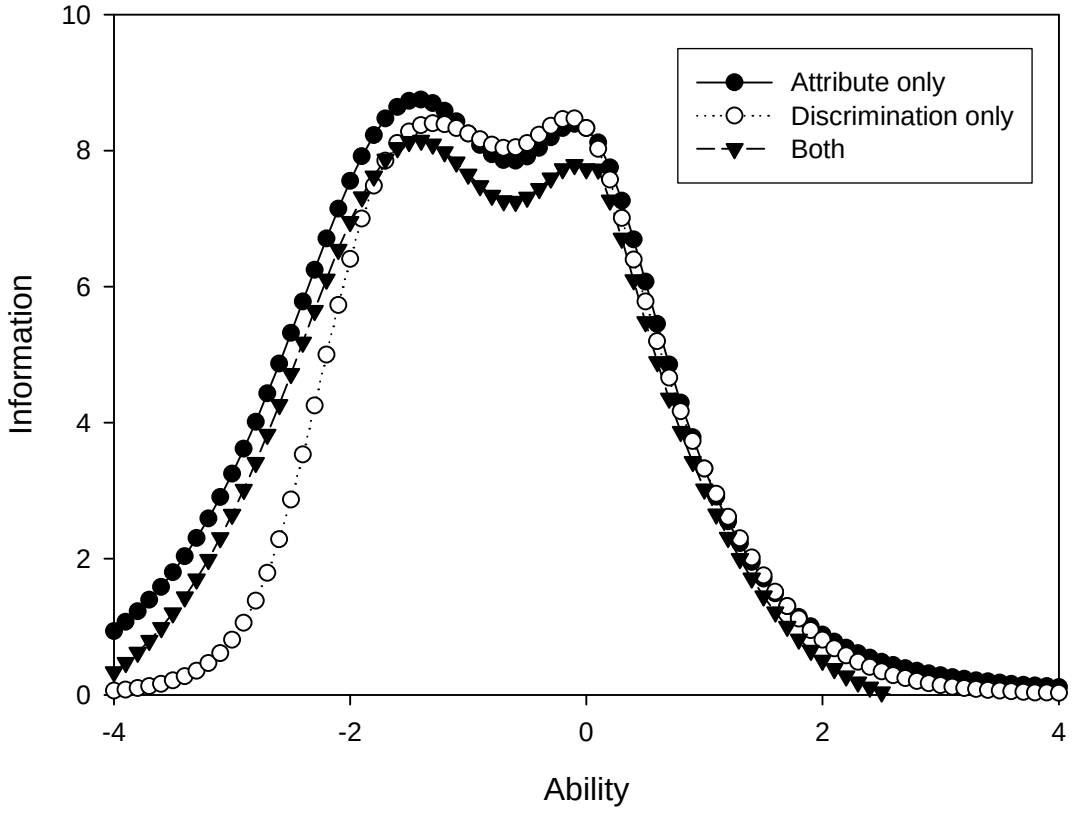
11. Express or solve problems using mathematical representation.

Appendix C

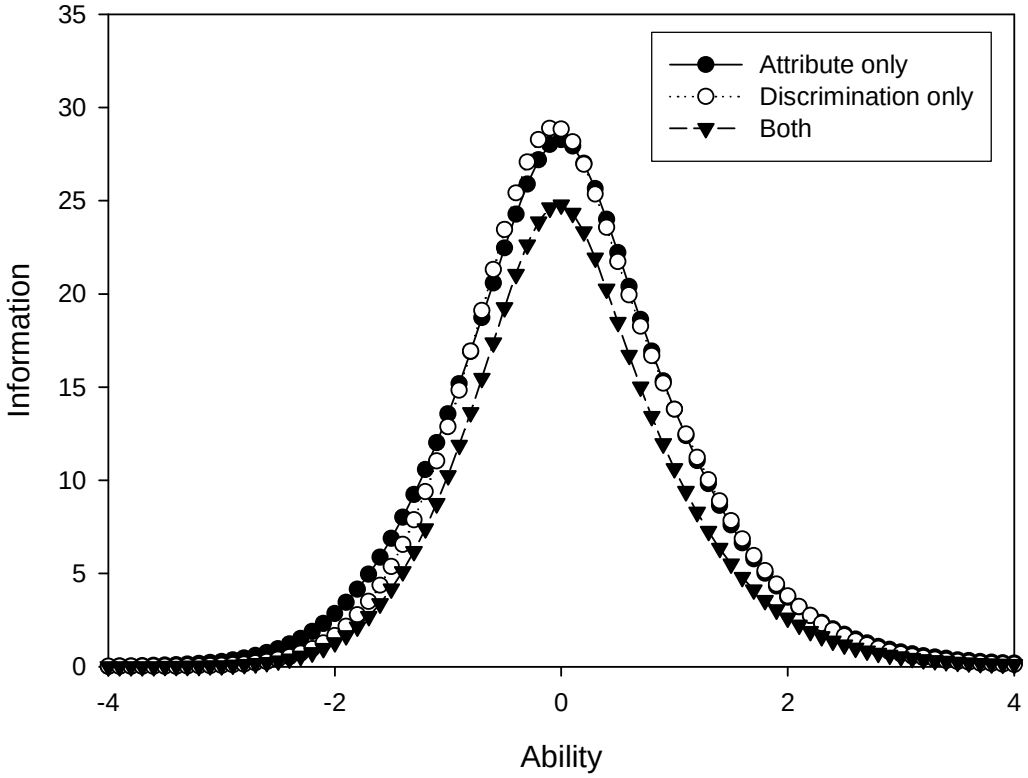
Test Information of three conditions

(Attribute constraint only, Discrimination only, and Both Attribute and Discrimination Constraint)

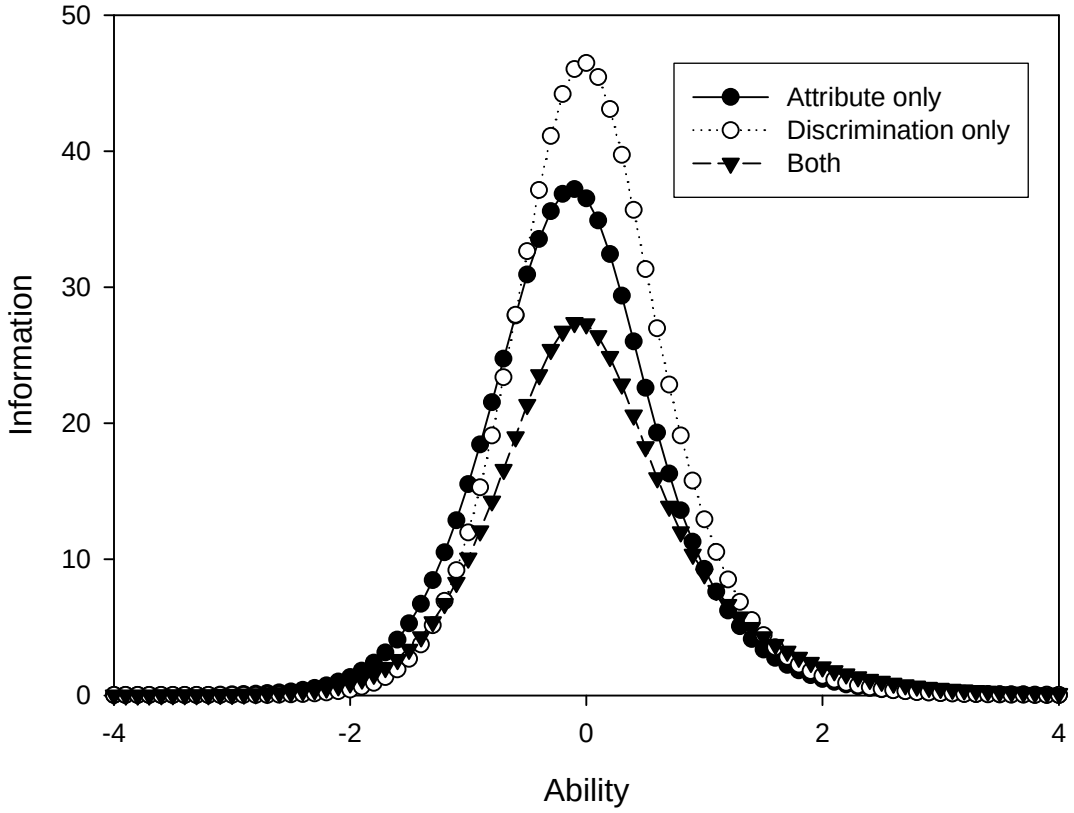
- Minimax Method of the first test.



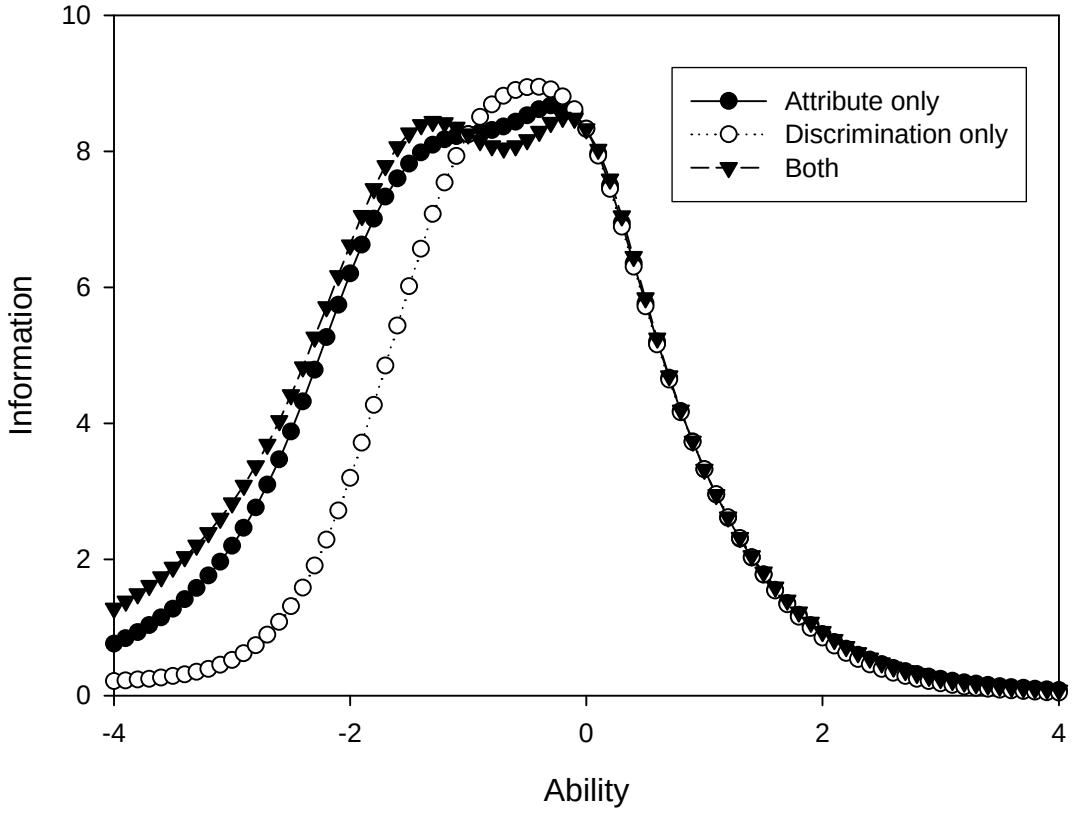
- Maximin Method of the first test



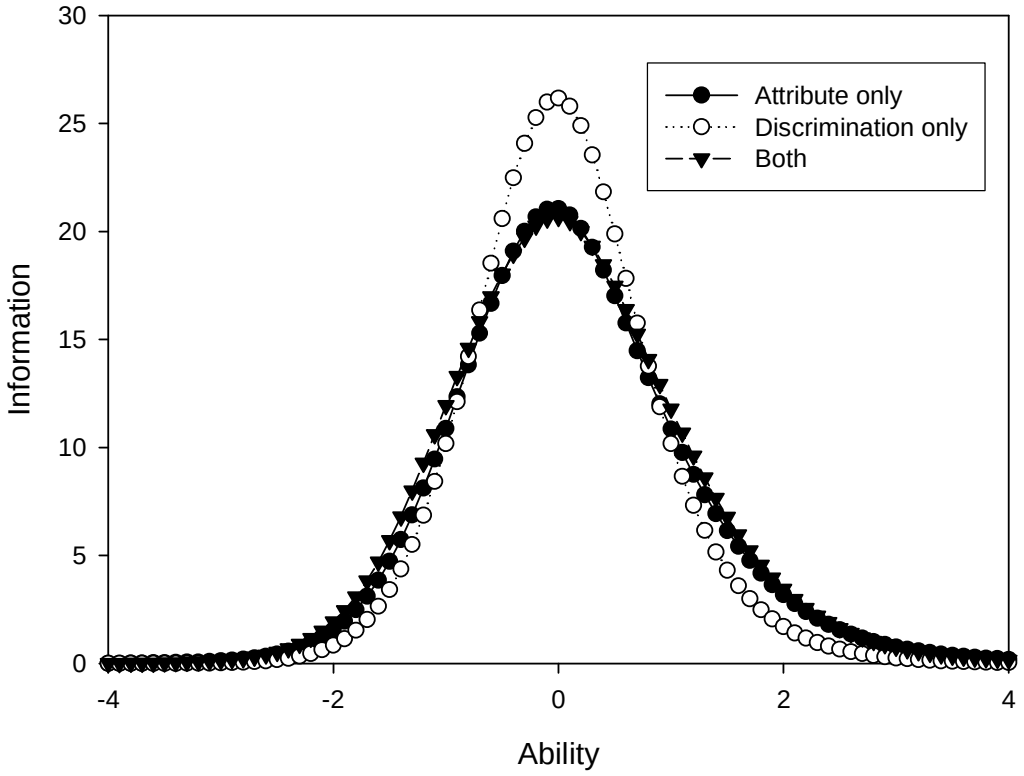
- Maximum Information Method of thefirst test



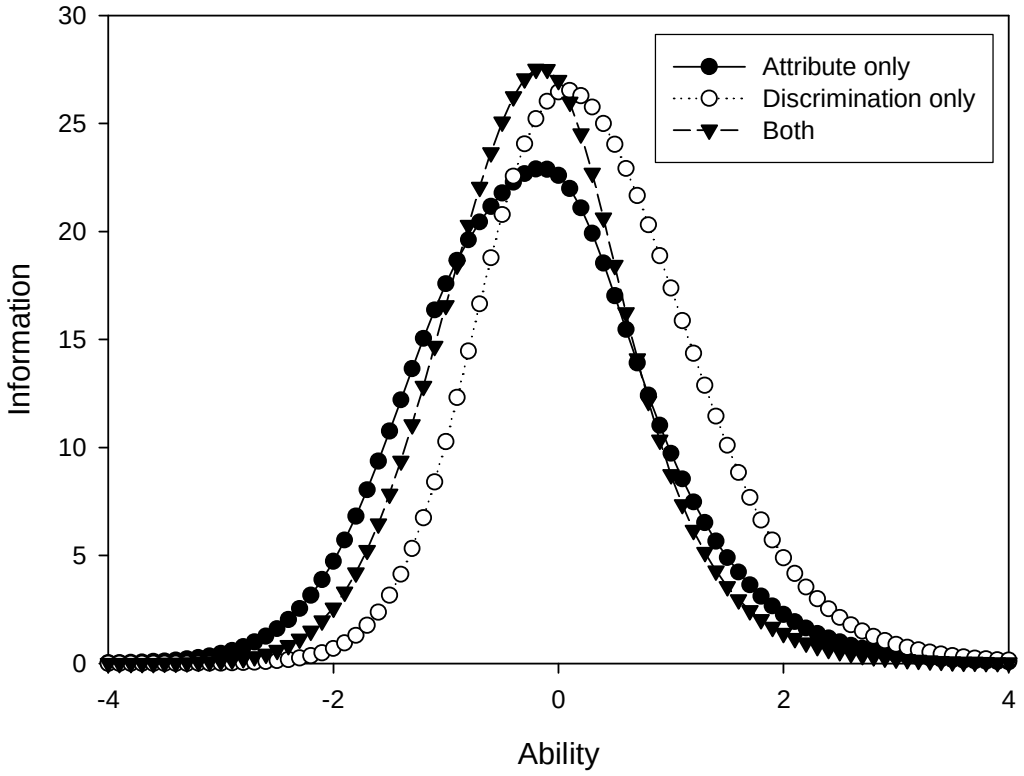
- Minimax Method of the second test



- Maximin Method of the second test



- Maximum Information Method of these cond test



Appendix D

The items selected and the Attributes selected for the Maximin Method, Attribute-only Condition

	Attributes										
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
56	0	1	0	0	0	0	0	0	0	0	0
57	0	0	0	1	0	0	0	0	0	0	0
65	1	0	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	1	0	0
81	0	0	0	0	0	0	1	0	0	0	0
84	0	0	0	0	0	0	0	1	0	0	0
107	0	0	1	0	0	0	0	0	0	0	0
108	0	0	0	0	1	0	0	0	0	0	0
119	0	0	0	0	0	0	0	0	0	0	1
120	0	0	0	0	0	0	0	0	0	1	0
124	0	0	0	0	0	1	0	0	0	0	0
188	0	1	0	0	0	0	0	0	0	0	0
189	0	0	0	1	0	0	0	0	0	0	0
197	1	0	0	0	0	0	0	0	0	0	0
199	0	0	0	0	0	0	0	0	1	0	0
213	0	0	0	0	0	0	1	0	0	0	0
216	0	0	0	0	0	0	0	1	0	0	0
239	0	0	1	0	0	0	0	0	0	0	0
240	0	0	0	0	1	0	0	0	0	0	0
241	0	0	0	1	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0	0	0	0	1
252	0	0	0	0	0	0	0	0	0	1	0
256	0	0	0	0	0	1	0	0	0	0	0
320	0	1	0	0	0	0	0	0	0	0	0
321	0	0	0	1	0	0	0	0	0	0	0
329	1	0	0	0	0	0	0	0	0	0	0
331	0	0	0	0	0	0	0	0	1	0	0
345	0	0	0	0	0	0	1	0	0	0	0
348	0	0	0	0	0	0	0	1	0	0	0
371	0	0	1	0	0	0	0	0	0	0	0
372	0	0	0	0	1	0	0	0	0	0	0
383	0	0	0	0	0	0	0	0	0	0	1
384	0	0	0	0	0	0	0	0	0	1	0

388	0	0	0	0	0	1	0	0	0	0	0
452	0	1	0	0	0	0	0	0	0	0	0
461	1	0	0	0	0	0	0	0	0	0	0
463	0	0	0	0	0	0	0	0	1	0	0
477	0	0	0	0	0	0	1	0	0	0	0
480	0	0	0	0	0	0	0	1	0	0	0
503	0	0	1	0	0	0	0	0	0	0	0
504	0	0	0	0	1	0	0	0	0	0	0
515	0	0	0	0	0	0	0	0	0	0	1
516	0	0	0	0	0	0	0	0	0	1	0
520	0	0	0	0	0	1	0	0	0	0	0
	4	4	4	4	4	4	4	4	4	4	4

The items selected and the Attributes selected for the Maximin Method,

Discrimination-only Condition

	Attributes										
Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
23	0	0	0	0	0	0	0	0	0	0	1
26	0	0	0	0	0	0	0	0	0	1	0
35	0	0	0	0	0	0	0	1	0	0	0
65	1	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	1	0
81	0	0	0	0	0	0	1	0	0	0	0
109	0	0	0	1	0	0	0	0	0	0	0
119	0	0	0	0	0	0	0	0	0	0	1
120	0	0	0	0	0	0	0	0	0	1	0
124	0	0	0	0	0	1	0	0	0	0	0
155	0	0	0	0	0	0	0	0	0	0	1
158	0	0	0	0	0	0	0	0	0	1	0
167	0	0	0	0	0	0	0	1	0	0	0
197	1	0	0	0	0	0	0	0	0	0	0
199	0	0	0	0	0	0	0	0	1	0	0
203	0	0	0	0	0	0	0	0	0	1	0
241	0	0	0	1	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0	0	0	0	1
252	0	0	0	0	0	0	0	0	0	1	0
256	0	0	0	0	0	1	0	0	0	0	0
287	0	0	0	0	0	0	0	0	0	0	1
290	0	0	0	0	0	0	0	0	0	1	0
299	0	0	0	0	0	0	0	1	0	0	0
329	1	0	0	0	0	0	0	0	0	0	0
335	0	0	0	0	0	0	0	0	0	1	0
373	0	0	0	1	0	0	0	0	0	0	0
383	0	0	0	0	0	0	0	0	0	0	1
384	0	0	0	0	0	0	0	0	0	1	0
388	0	0	0	0	0	1	0	0	0	0	0
419	0	0	0	0	0	0	0	0	0	0	1
422	0	0	0	0	0	0	0	0	0	1	0
431	0	0	0	0	0	0	0	1	0	0	0
461	1	0	0	0	0	0	0	0	0	0	0
467	0	0	0	0	0	0	0	0	0	1	0
477	0	0	0	0	0	0	1	0	0	0	0
505	0	0	0	1	0	0	0	0	0	0	0
515	0	0	0	0	0	0	0	0	0	0	1
516	0	0	0	0	0	0	0	0	0	1	0
	4	0	0	4	0	3	2	4	1	12	8

The items selected and the Attributes selected for the Maximin Method, Both Attribute and Discrimination Condition

	Attributes										
Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
56	0	1	0	0	0	0	0	0	0	0	0
65	1	0	0	0	0	0	0	0	0	0	0
73	0	0	0	0	0	0	0	0	1	0	0
81	0	0	0	0	0	0	1	0	0	0	0
100	0	0	1	0	0	0	0	0	0	0	0
108	0	0	0	0	1	0	0	0	0	0	0
109	0	0	0	1	0	0	0	0	0	0	0
119	0	0	0	0	0	0	0	0	0	0	1
167	0	0	0	0	0	0	0	1	0	0	0
188	0	1	0	0	0	0	0	0	0	0	0
197	1	0	0	0	0	0	0	0	0	0	0
205	0	0	0	0	0	0	0	0	1	0	0
213	0	0	0	0	0	0	1	0	0	0	0
232	0	0	1	0	0	0	0	0	0	0	0
240	0	0	0	0	1	0	0	0	0	0	0
241	0	0	0	1	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0	0	0	0	1
252	0	0	0	0	0	0	0	0	0	1	0
256	0	0	0	0	0	1	0	0	0	0	0
299	0	0	0	0	0	0	0	1	0	0	0
372	0	0	0	0	1	0	0	0	0	0	0
383	0	0	0	0	0	0	0	0	0	0	1
384	0	0	0	0	0	0	0	0	0	1	0
388	0	0	0	0	0	1	0	0	0	0	0
431	0	0	0	0	0	0	0	1	0	0	0
452	0	1	0	0	0	0	0	0	0	0	0
461	1	0	0	0	0	0	0	0	0	0	0
469	0	0	0	0	0	0	0	0	1	0	0
477	0	0	0	0	0	0	1	0	0	0	0
496	0	0	1	0	0	0	0	0	0	0	0
505	0	0	0	1	0	0	0	0	0	0	0
515	0	0	0	0	0	0	0	0	0	0	1
516	0	0	0	0	0	0	0	0	0	1	0
520	0	0	0	0	0	1	0	0	0	0	0
	3	3	3	3	3	3	3	3	3	3	4

Bibliography

Adema, J. J. (1992). Methods and models for the construction of Weakly Parallel Tests. *Applied Measurement in Education 16*, 53-63.

Adema, J. J., & van der Linden, W. J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics, 14*, 279-290.

Baker, F. B., Cohen, A. S., & Barmish, B. R. (1988). Item characteristics of tests constructed by linear programming. *Applied Psychological Measurement, 12*, 189-200.

Bertsimas, D. & Tsitsiklis, J. N. (1997). Introduction to Linear Optimization. Belmont, Massachusetts: Athena Scientific.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores. Reading*, MA: Addison-Wesley Publishing Company.

Birenbaum, M. and Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education, 6*, 225-268.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1*, 225-268.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15*, 129-145.

Boekkooi-Timminga, E., & van der Linden, W.J. (1987). Algorithms for automated test construction. In F.J.Maarse, L.J.M. Mulder, W.P.B. Sjouw, & A.E. Akkerman, *Computers in psychology: Methods, instrumentation and psychodiagnostics* (pp. 165-170). Lisse: Swets & Zeitlinger.

Boisvert, R.F., Howe, S.E., and Kahaner, D.K. (1985) GAMS: A Framework for the Management of Scientific Software, *ACM Transactions on Mathematical Software* 11, 4.

Chang, H. & Ying, Z. (1996). A global information approach to computerized adaptive testing, *Applied Psychological Measurement*, 20, 213-229.

Chang, H. & Ying, Z. (1999). α-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-221.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 1-18). Hillsdale, NJ: Lawrence Erlbaum Associates.

de Gruijter, D. N. M. (1990). Test construction by means of linear programming. *Applied Measurement in Education, 14*, 175-181.

DiBello, L., Stout, W., & Rousses, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 361-389). Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica 37*, 359-374.

Fletcher, R. B. (2003). Automated test assembly using 0-1 linear programming. A didactic explanation.

Gass, S.I. (1985). *Linear programming: methods and applications*, New York, NY: McGraw-Hill.

Hambleton, R. K., & Swaminathan, H. (1985). *Item respsonse theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hartz, S. (2002). A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice. Doctoral thesis, The University of Illinois at Urbana-Champaign.

Hartz, S., Roussos, L., and Stout, W. (2002) *Skills Diagnosis: Theory and Practice*. User Manual for Arpeggio software. ETS.

Hawkins, D. M. (1988). Branch-and-bound method. In S. Kotz, N. L. Johnson and C. B. Read, Eds.), *Encyclopedia of Statistical Sciences, Vol 1*. (p. 314-316). New York, NY: John Wiley and Sons.

ILOG, Incorporation. (2003). CPLEX Software Program, version 8.1. Incline Village, NV: CPLEX Division.

Jensen, P.A. & Bard, J.F. (2003). *Operations Research Models and Methods*, New York, NY: John Wiley and Sons.

Jiang, H. (1996). Applications of Computational Statistics in Cognitive Diagnosis and IRT Modeling. Doctoral thesis, The University of Illinois at Urbana-Champaign.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*, 149-174.

McGlohen, M. (2004). The Application of Cognitive Diagnosis and Computerized Adaptive Testing to a Large-Scale Assessment. Doctoral thesis, The University of Texas at Austin.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Nemhauser, G., & Wolsey, L. (1988). *Integer and combinatorial optimization*. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests,* Denmarks Pedagogiske Institut, Copenhagen.

Rogers, H. J., Swaminathan, H. and Hambleton, R. K. (1991). *Fundamentals of item response theory: Measurement methods for the social sciences volume 2.* Thousand Oaks, CA: Sage Publications.

Samejima, F. (1995). A cognitive diagnosis model using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive psychometric diagnosis model. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 391-410). Hillsdale, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1997). Graded response model. In W.J. van der Linden and R. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.

Swanson, L., & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20*(4).

Tatsuoka, K. K. (1984) Caution indices based on item response theory. *Psychometrika 49*(1), 95-110.

Tatsuoka, K. K. (1990). Toward integration of item response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p.453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. and Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics 7*(3), 215-231.

Tatsuoka, K. K. and Tatsuoka, M. M. (1984). Bug distribution and pattern classification. *Psychometrika 52*(2), 193-206.

Tatsuoka M. M. and Tatsuoka, K. K. (1989). Rule space. In S. Kots and N. L. Johnson (Eds.) Encyclopedia of statistical sciences (vol. 8, pp. 217-220). New York: Wiley.

Texas Education Agency (2002). *Texas Student Assessment Program Technical Digest for the Academic Year 2001-2002*, Austin, TX.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika 50*, 411-420.

Theunissen, T. J. J. M. (1986). Optimization algorithms in test design. *Applied Psychological Measurement, 10*, 381-389.

Timminga, E., & Adema, J. J. (1995). Test construction from item banks (pp. 111-127). In G. H. Fischer & I. W. Molenaar (Eds.), *The Rasch model: Foundations, recent developments, and applications.* New York: Springer-Verlag.

U.S. House of Representatives (2001). Text of the 'No Child Left Behind Act'. Public Law No. 107-110, 115 Stat. 1425.

van der Linden, W. J. (1994). Optimum design in item response theory: Applications to test assembly and item calibration. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 308-318). New York: Springer-Verlag.

van der Linden, W. J. (1987). Automated test construction using minimax programming. In W. J. van der Linden (Ed.). *IRT-based test construction* (Research Report 87-2, chap. 3). Enschede: University of Twente, Department of Education.

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20*, 373-388.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement, 24*, 225-240.

van der Linden, W. J. (in press). *Linear Models for Optimal Test Design*. New York: Springer-Verlag.

van der Linden, W. J. & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. *Applied Psychological Measurement, 12*, 201-209.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika 54*(2), 237-247.

van der Linden, W. J., & Luecht, R. M. (1996). An optimization model for test assembly to match observed-score distributions. In G. Engelhard Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 405-418). Norwood NJ: Ablex.

van der Linden, W. J., & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259-270.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (p. 61-100). Mahwah, NJ: Lawrence Earlbaum Associates.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

Vita

Soojin Kim was born in Seoul, Korea on May 21, 1976, the daughter of Dae Hee Kim and Young-Ha Kim. After completing her work at Sacred Heart Girls' High School, Seoul, Korea in 1995, she entered Chung-Ang University in Seoul, Korea. She received the degree of Bachelor of Arts from the Chung-Ang University in February 1999. In September 1999, she entered the Graduate School at the University of Texas at Austin to work on a Master's in Educational Psychology. While attending the Graduate School, she worked as a teaching assistant at the Educational Psychology department and a research assistant at the Measurement and Evaluation Center (MEC). After receiving the master's degree, she continued to her studying in Ph.D program in Educational Psychology at the University of Texas at Austin.

Permanent address: 38-248, Yongmoon-dong, Youngsan-gu, Seoul, Korea

This dissertation was typed by the author.